Supporting Information File

QSRR Models to Support Non-target High Resolution Mass Spectrometric Screening of Emerging Contaminants in Environmental Samples

Reza Aalizadeh, Nikolaos S. Thomaidis, Anna A. Bletsou and Pablo Gago-Ferrero*

Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis Zografou, 15771 Athens, Greece

*****Corresponding author.

Tel.:+302107274317 – Fax: +302107274750

E-mail address: ntho@chem.uoa.gr

For submission to: Journal of Chemical Information and Modeling

Table of contents:

SI 1. Supplementary content to Material and Methods section

SI 1.1. Chemicals

The reference standards of the pesticides were donated to the laboratory by Bruker Daltonics (Bremen, Germany), at a concentration of 1 mgL-1 in methanol. The rest of the compounds included in the study were all purchased from Sigma–Aldrich (Germany)**.** Individual stock solutions of these compounds were prepared in methanol at a concentration of 1 g L^{-1} and stored at -20 °C. Then, working solutions were prepared in methanol at a concentration of 1 mg L^{-1}. Methanol, LC-MS grade, was purchased from Merck (Germany), whereas 2-propanol of LC-MS grade was from Fisher Scientific (Geel, Belgium). Sodium hydroxide monohydrate (NaOH) for trace analysis ≥99.9995%, ammonium acetate, ammonium formate and formic acid, all LC-MS grade, were purchased from Fluka, Sigma–Aldrich (Germany). Distilled water used for LC–MS analysis was provided by a Milli-Q purification apparatus (Millipore Direct-Q UV, Bedford, MA, USA). Regenerated cellulose (RC) syringe filters (15 mm diameter, 0.22 μm pore size) were provided from Phenomenex (Torrance, CA, USA).

SI 1.2. Instrumentation

An ultrahigh-performance liquid chromatography (UHPLC) system with a LPG-3400 pump (Dionex UltiMate 3000 RSLC, Thermo Fisher Scientific, Germany), interfaced to a QToF mass spectrometer (Maxis Impact, Bruker Daltonics, Bremen, Germany) was used for the screening analysis.

The chromatographic separation was performed on an Acclaim RSLC C18 column (2.1 × 100 mm, 2.2 μm) from Thermo Fisher Scientific (Driesch, Germany) preceded by a guard column, ACQUITY UPLC BEH C18 1.7 μm, VanGuard Pre-Column, Waters (Ireland), thermostated at 30 ˚C. Mobile phase composition in positive ionization mode (PI) is (A) H_2O:MeOH (90:10) with 5 mM ammonium formate and 0.01% formic acid and (B) MeOH with 5 mM ammonium formate and 0.01% formic acid. For the negative ionization mode (NI), the mobile phase is (A) H_2O:MeOH (90:10) with 5 mM ammonium acetate and (B) MeOH with 5 mM ammonium acetate.

The gradient elution program was the same for the two ionization modes and the chromatogram lasts 15.5 min, with 5 min of re-equilibration of the column for the next injection. It starts with 1% B with a flow rate of 0.2 mL min^{-1} for 1 min and it increases to 39 % in 2 min (flow rate 0.2 mL min^{-1}), and then to 99.9 % (flow rate 0.4 $mLmin^{-1}$) in the following 11 min. Then it keeps constant for 2 min (flow rate 0.48 mL min^{-1}) and then initial conditions were restored within 0.1 min and the flow rate decreased to 0.2 mL min^{-1}. The injection volume was set up to 5 μL.

The operating parameters of the electrospray ionization interface (ESI) are for PI mode: capillary voltage, 2500 V; end plate offset, 500 V; nebulizer, 2 bar; drying gas, 8 L min^{-1}; dry temperature, 200 °C; and for NI mode: capillary voltage, 3500 V; end plate offset, 500 V; nebulizer, 2 bar; drying gas, 8 L min^{-1}; dry temperature, 200 °C.

The QToF MS system operates in broadband collision induced dissociation (bbCID) acquisition mode and records spectra over the range m/z 50−1000 with a scan rate of 2 Hz. The Bruker bbCID mode provides MS and MS/MS spectra at the same time, while it works at two different collision energies. At low collision energy (4 eV), MS spectra were acquired and at high collision energy (25 eV), fragmentation is taking place at the collision cell resulting in MS/MS spectra.

A QToF external calibration was daily performed with a sodium formate solution, and a segment (0.1−0.25 min) in every chromatogram was used for internal calibration, using a calibrant injection at the beginning of each run. The sodium formate calibration mixture consists of 10 mM sodium formate in a mixture of water is opropanol (1:1). The theoretical exact masses of calibration ions with formulas $Na(NaCOOH)_{1-14}$ in the range of 50−1000 Da were used for calibration. The instrument provided a typical resolving power of 36000−40000 during calibration (39274 at m/z 226.1593, 36923 at m/z 430.9137, and 36274 at m/z 702.8636). Mass spectra acquisition and data analysis was processed with Data Analysis 4.1 and Target Analysis 1.3 (Bruker Daltonics, Bremen, Germany).

SI 1.3. Sampling details and sample preparation

The sample used in the collaborative trial was collected from location JDS57, downstream of Ruse/Giurgiu (RO/BG; rkm 488; coordinates N43.890150, E26.017067) on September 18, 2013 as a part of the Third Joint Danube Survey, organized by the International Commission for the Protection of the Danube River (ICPDR). The sample preparation included a large-volume solid-phase extraction (LVSPE) of 1000 litres of water. Briefly, the sampler cartridge was filled with 160 g of Macherey Nagel Chromabond® HR-X (neutral resin) and 100 g each of Chromabond® HR-XAW (anionic) and HR-XCW (cationic exchange resin). The resins were extracted with 500 mL each of ethyl acetate and methanol (HR-X), 500 mL methanol with 2% 7 M ammonia in methanol (HR-XAW) or 500 mL methanol with 1% formic acid (HR-XCW). The extracts were then combined, neutralized, filtered (Whatman GF/F) and reduced to a final volume of 1 L using rotary evaporation. Aliquots of 1.5 mL, equivalent to 1.5 L of river water, were transferred into vials and evaporated to dryness under nitrogen. These were sent to each participant along with a laboratory blank. The samples were reconstituted in MeOH:H20 (50:50) in 1.5 mL and filtered through RC syringe filters prior to analysis.

SI 2. Supplementary content to Section 2.5: Division of the datasets

Dividing the data set into training and test sets is one of the crucial steps in QSPR modeling for creating internally and externally appropriate predictive models. Therefore, despite random selection of the test set, some techniques should be aplied for preventing information to be lost. There are different methods including k-means, k-nearest neighborhood, principle component analysis and selection based on distribution of properties. In this work, principle component analysis (PCA)[1] was carried out to compress the molecular descriptors into principle components (PCs).[2] The selection of meaningful PCs is the major problem in an analysis based on PCA. The common method is top-down variable selection that PCs remarked as order of decreased eigenvalues, and then, the highest eigenvalue of PCs is chosen as the most significant PCs. In this work, two PCs have been obtained including PC1 and PC2 with 14.85% and 9.78% contribution to the total PCs in negative ionization, and created the total value of 24.63% of the variation in the whole data. For the positive ionization, the same analysis was used and showed PC1 and PC2 with 16.38% and 9.84% contribution to the total PCs of 27.21%, respectively. The results shown in **Figure S1 (A) and (B)** demonstrate the scatter distribution of data for each one of PCs. As it can be seen, training and test set scatter evenly in 2D space in which confirming that this method can be employed to split the data set. Since both kNN and PCA are given based on the descriptors obtained after removal collinear variables, they are good indicators of molecular diversity and similarity. The presentation of kNN classification is much better due to the generation of graphical analysis which demonstrates the similarity distance and clustering group for each compound. In this technique, each molecule based on its molecular descriptors is classifying in a cluster, and then another compound with near similarly is being searched and classified in the same cluster with the previous case. Therefore, the analysis is started with a separated cluster, and then continues with combinations of them, and finally ends when there is only one cluster left.[3,4] The results of the analysis are given as dendrograms, where the linkage distance with information about cluster population can be observed. The classification results in negative and positive ionization are given in **Figure S1 (C) and (D)**, respectively. The selection of test set and training set is based on two principles: (1) the selection should cover from the lowest to the highest retention

times, and (2) based on the population of each cluster, the respective compounds with the largest linkage distance from previous selection should be considered. For each model in subsequent analysis, the selected test set compounds (with the ratio of 20% of total molecules) are shown in Table S1.

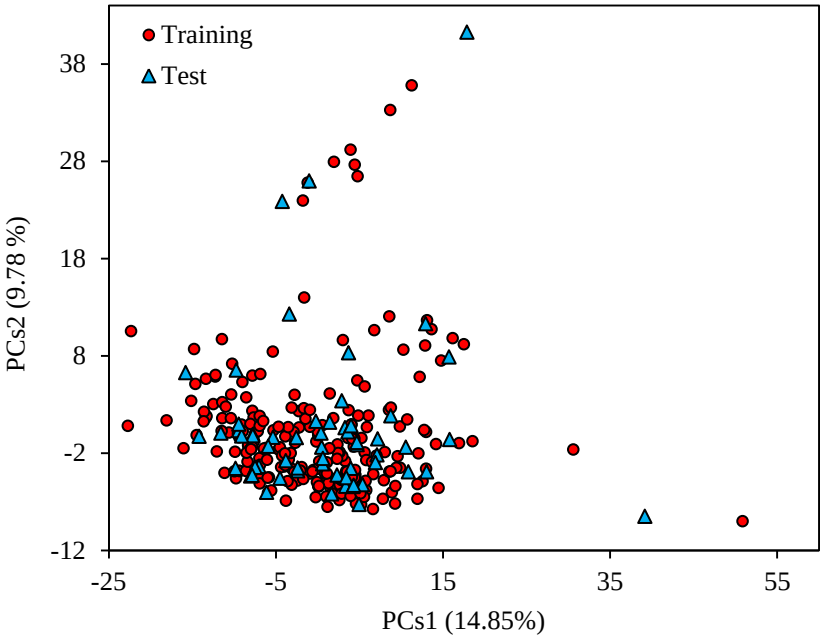


Figure S1 (A). PCA analysis for negative ionization compounds (sample test set for SW-MLR)

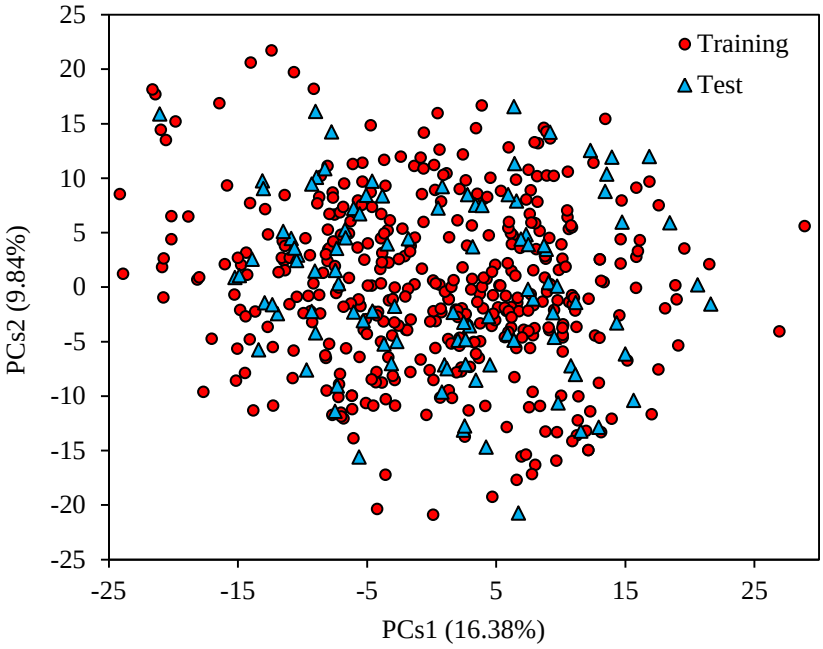


Figure S1 (B). PCA analysis for positive ionization (sample set for SW-MLR)

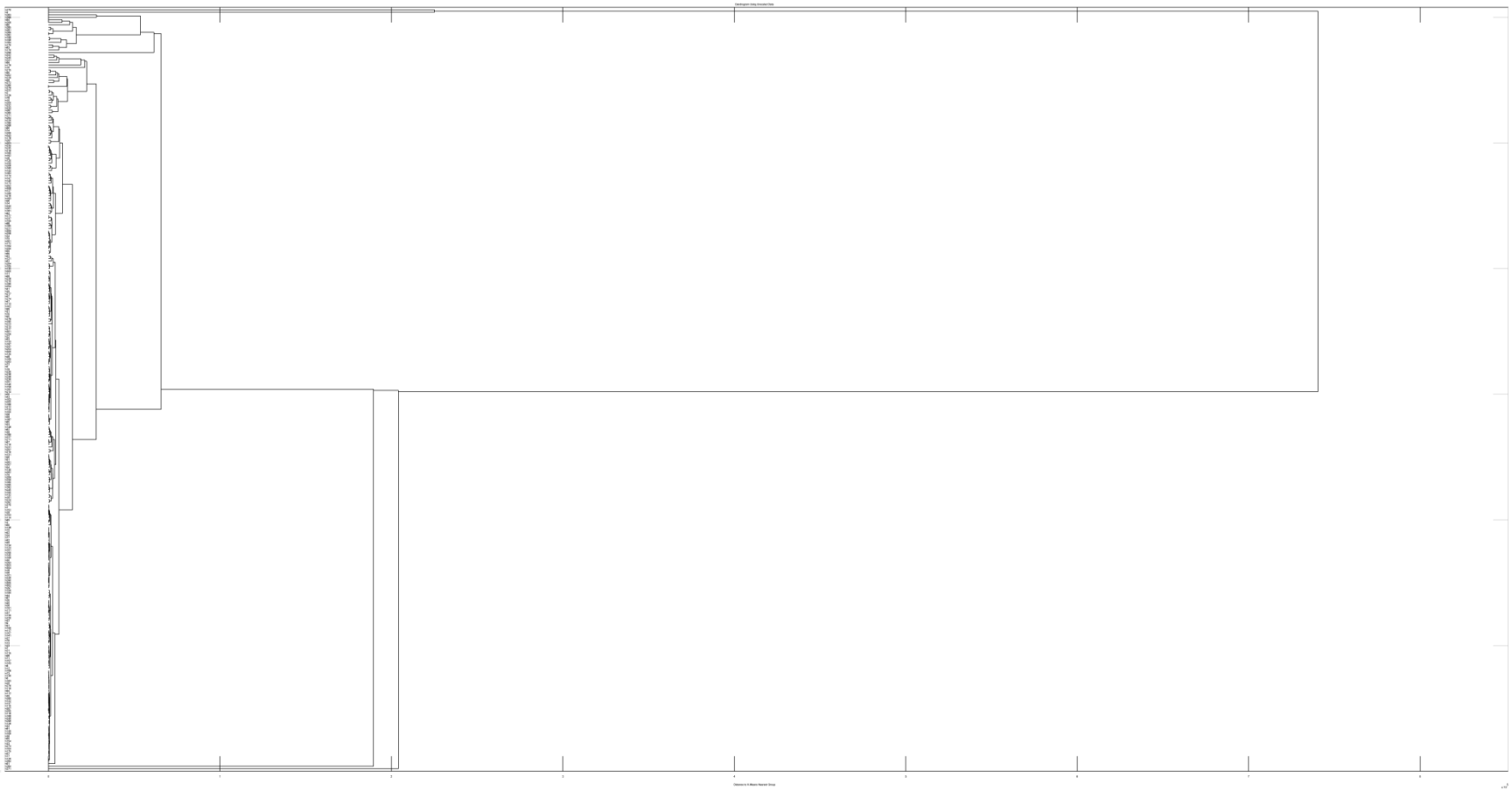


Figure S1 (C). kNN analysis for negative ionization compounds

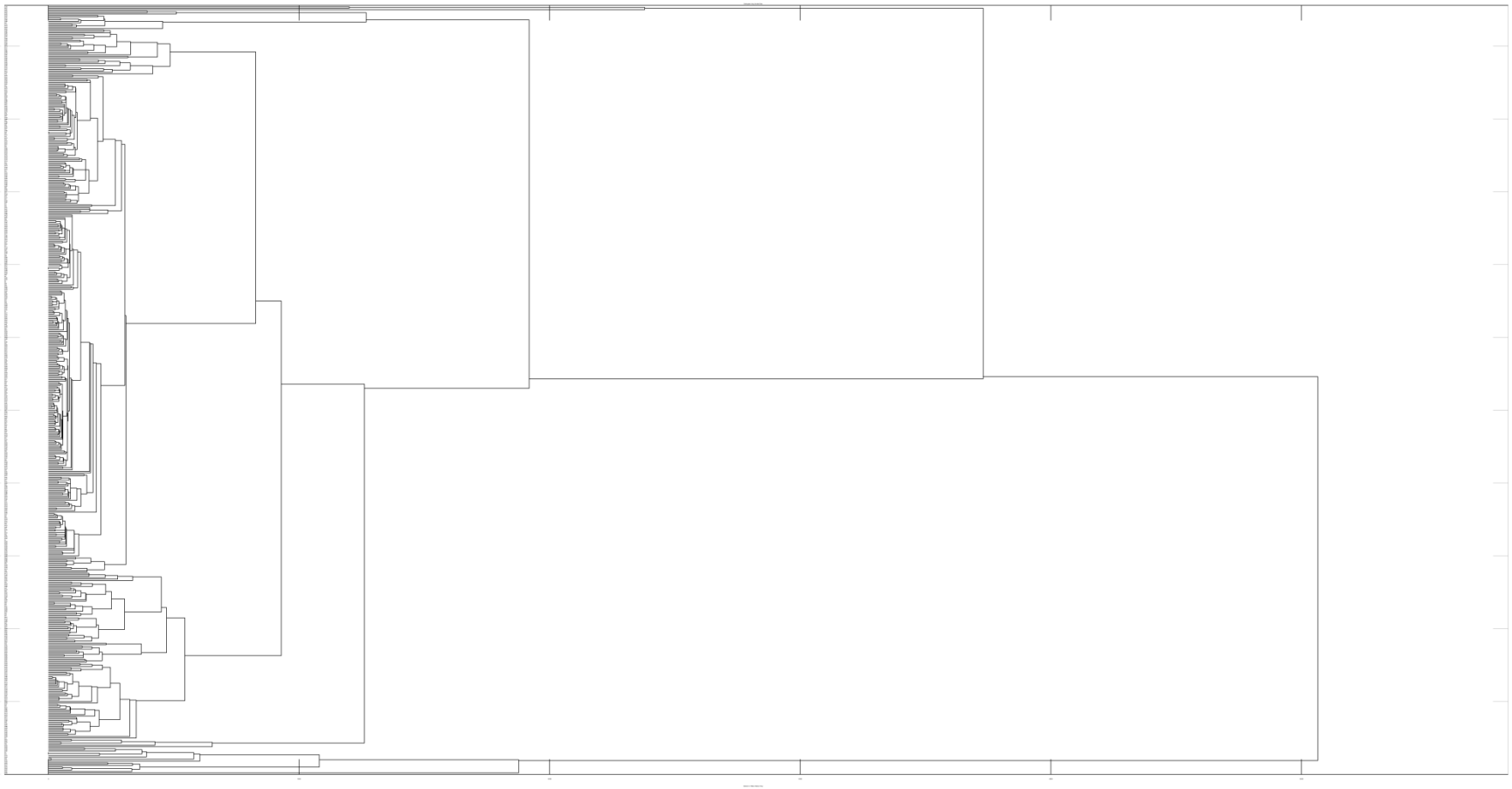


Figure S1 (D). kNN analysis for positive ionization compounds

SI 3. Supplementary content to Section 2.5: Variable selection methods

Stepwise selection algorithm

Stepwise selection technique is a well-known and simple method for identifying the correct number of variables in data matrix when the procedure includes a regression models for its selection base.[5, 6] The stepwise variable selection technique was applied by forward selection and back elimination rule where the variable possessed the highest correlation value with retention time (experimental data) is being selected, and based on the regression model, its regression coefficient is being calculated. Each selected variable (here the molecular descriptor) is then tested using F-test[6-9] to evaluate its significance and contribution to the model. If it improves the model, it will be retained in the model. This procedure is called forward selection. In addition, all variables were selected in backward elimination steps[5, 8] and if the selected variable did not contribute in improvement of the model, it was excluded from the set of significant variables and was eliminated from the model. The two steps were continued until no further improvement was observed. The only disadvantage of this technique was the over-fitting, since the selection is based on data fitting; to prevent this problem, a cross-validation method should be employed to evaluate the predictive ability of the proposed model.[7-9]

Genetic Algorithm

Genetic algorithm (GA) is a global optimization method and is based on the evolutionary computations. It finds solutions for existed problems on the biases of survival of the fittest rule.[10] The GA technique starts with binary coding of molecular descriptors values (initial population of objects) for each compound to permit the mathematical treatment of "chromosomes". "Chromosomes" are randomly selected group of molecular descriptors that represent each individual subsets of population.[11] The descriptors inside these "chromosomes" are called "genes".[12, 13] The state of variables selected is shown by value "1" and the non-selected variable represented by "0" value.[14] To collect a small subset, and for the simplicity in the interpretation of results, generally the number of selected variables is kept proportionally. Therefore, the population of non-selected descriptors (indicated by zero, with the ratio of 60%) is more than that presented by selected variables.[11, 12, 14] The total number of "chromosomes" is indicating the

population (generally lies between 50 and 500) which is depending on the dimension of the problem. These "chromosomes" are evaluated based on the *fitness function* (here is the correlation coefficient of leave-one-out cross validation (Q^2_{LOO})), so that if chromosomes couldn't meet the cut-off criteria, they are being stopped from spreading to the next generations. Next, the survived "chromosomes" are reproducing new number of population, and the probability level of each "chromosome" is calculated based on its outcomes associated with the taken responses. The best number of "chromosomes" would be finally selected by their higher probability which results in better response. The cross-over technique is then being applied to these "chromosomes" to pair them in a new generation for deriving the most effective "genes" in "chromosomes". Finally, mutation, which causes to impose values that are not tried for each descriptor, is being applied to newly derived generation.[15] The reproduction and mutation of "chromosomes" continue until the best number of descriptors in a "chromosome" is selected within the GAs iteration of generation. The disadvantages of this selection tool are its tendency to select both relevant and irrelevant descriptors for regression and generation of first population that can be resolved by repeating the Genetic Algorithm calculation.[10, 16] Therefore, the obtained results can introduce the most significant variables for constructing predictive model.[10] To get the best combination of descriptors set as the variables for subsequent analysis, the different subset of descriptors based on genetic algorithm were obtained and the models were built for each set.

SI 4. Supplementary content to Section 2.5: Modeling chemometric techniques

Multiple linear regressions

Multiple linear regressions (MLR) is one of the most used linear method in QSRR. To derive a MLR model, the number of molecules in data set should be five times higher than the number of selected descriptors (the descriptors should be orthogonal). A low number of descriptors are of interest in order to minimize the information overlap in descriptors. In this work, to obtain the best linear model, the statistical parameters (R^2 and Q^2 values) were considered. The MLR model provided a linear equation which is linking the structural features to the retention times of the compound:

$$Rt = a_0 + b_1x_1 + \cdots + b_nx_n \qquad (Eq.S1)$$

where a_0 is the intercept and the b_i is regression coefficients of the selected descriptors x_i . To evaluate the strengths and goodness of the model, the coefficient of multiple determinations was used. R^2 value calculates the proportion of the variation in the response and was obtained as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{I}(y_i - \bar{y})^2} \qquad (Eq.S2)$$

where y_i is the observed property/activity (here is the experimental retention time), $\bar{y}$ the mean value of the experimental data and $\hat{y}_i$ the calculated retention time. The R^2 value higher than 0.5 and near 1.0 indicates the acceptable predictive ability of the model. Generally, the R^2 value can change (either increased or decreased) by adding extra variables to the model. Therefore, this problem can be solved considering the adjusted R^2 values (R^2_{adj}):

$$R^2_{adj} = \left[1 - (I - R^2)\left(\frac{I-1}{I-n-1}\right)\right]^{1/2} \qquad (Eq.S3)$$

In this equation, (*I*) is the number of calibration objects, and (*n*) is the number of the selected descriptors for model. The statistical significance of the proposed model can also be given by the null hypothesis where this implies that all these descriptors in the model beyond the constant value are required for modeling. To derive the given null hypothesis, the F-value can be used as follows:

$$F = \frac{(n-k-1)\sum_{i=1}^{n}(y_i - \bar{y})^2}{k\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (Eq.S4)$$

where the *n* is the number of the compounds in the data set and *k* is the number of descriptors. The higher the F –value becomes, the greater the probability that the equation is significant. Therefore, the procedure results in selection of the appropriate and relevant descriptors if its null hypothesis is rejected by having the higher F values. Another important statistical parameter that is used in both linear and non-linear methods to validate the outcome of the derived models is the root mean square error (RMSE), where the lower RMSE value indicates the less error generated by the built models, and thus, the model can be accepted for prediction purposes. The RMSE value is calculated as below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (Eq.S5)$$

The most important statistical parameter used in the validation of MLR models is the cross-validation correlation coefficient, which is calculated as leave-one-out compound principle. In every calculation process for obtaining Q^2_{LOO} value, one of the compounds in the dataset is being excluded from the model and its activity is calculated from the proposed model. This process is continued until all available compounds in the data matrix are excluded once, and their activities are being predicted by the model. Therefore, this technique is a good indicator of the strength of the derived models. A robust model should implement higher Q^2_{LOO} value. This value can be calculated as follows:

$$Q^2_{LOO} = r^2_{cv} = 1 - \frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{I}(y_i - \bar{y})^2} \qquad (Eq.S6)$$

Further, the external predictive ability of the constructed model can be assessed by modified r^2 value (Eq. S7) and the concordance correlation coefficient (Eq. S8), evaluating both accuracy and precision.[17] Concordance correlation coefficient (CCC) evaluates the degree to which pairs of observations fall on the 45° line through the origin.

$$r^2_m = r^2\left(1 - \left|\sqrt{r^2 - r^2_o}\right|\right) \qquad (Eq.S7)$$

where r^2 and r^2_o are squared correlation coefficients between the observed and predicted retention time value of the test set compounds with and without intercept, namely.

$$CCC = \rho C_b \qquad (Eq.S8)$$

where ρ is the Pearson correlation coefficient, and measures how far each observation deviates from the best-fit line. Thus the ρ value is a measure of the precision, and C_b is the bias correction factor *i* which calculates how far the best-fit line deviates from the 45° line through the origin, and therefore, it is a measure of the accuracy.

Support vector machines

Support vector machines are used for building nonlinear models. This technique adopts the structure risk minimization (SRM) principle by introducing <*epsilon*> (ε) *insensitive loss function* to solve the problems. This constitutes a trade-off between the

complexity of the model and its capability to reproduce experimental observations. The regression outcome depends highly on a good setting of parameters: C, ε, the kernel type, and corresponding kernel parameters. Parameter C is a regularization constant which calculates the trade-off between the model complexity and the degree in which deviations larger than ε are tolerated in optimization formulation. The selection of the kernel function and corresponding parameters are considered one of the crucial steps in performing SVM modeling, because they define the distribution of the training set of samples in the high dimensional feature. In this work, the radial basis function (RBF) was used as kernel function where the most significant parameter is the width (γ) of the radial basis function. All calculations regarding implementation of SVM and its analyses were written in MATLAB package.[18] The radial basis function (RBF) is defined as follows:

$$k(\bar{x}_i, \bar{x}_j) = \exp\left(-\gamma \|\bar{x}_i - \bar{x}_j\|^2\right) \qquad (Eq.S9)$$

where k refers to the kernel function and γ is a parameter of kernel, $\bar{x}_i$ and $\bar{x}_j$ are independent variables.

Artificial neural network modeling

In this work, the feed-forward artificial neural networks with back-propagation of error algorithm was employed to derive the ANN nonlinear models. The input for the model generation is the variables (descriptors) selected by genetic algorithm. The initial weights were randomly chosen between 0 and 1.[19] Optimization of the weights and biases is performed based on the resilient back-propagation algorithm. The complex step in performing the ANN model is the identification of the correct hidden layers to generate the QSRR model.[19] Therefore, a three-layer network with a sigmoidal transfer function was designed.[20] To obtain the correct nodes in the hidden layers, RMSE values were considered for both test and training sets, and the nodes with the lower RMSE is selected for subsequent analysis.[19] In this work, the designed network was accomplished and performed in 20000 iterations. However, in most cases,[21] increasing the iterations would increase the value of standard error of prediction and therefore, over-fitting occurs.[21] The data set was divided into three groups using principle component analysis and clustering techniques, separately: a training set, a validation set and a prediction set for negative and positive ionization are consisting of 181:58:59, and 315:105:105 molecules, respectively.

The training and validation sets were employed for deriving the predictive models and the prediction set was used for evaluating the external abilities of the generated models.[21] For obtaining the best model, despite RMSE, R^2 and mean percentage deviation (MPD) values for the results of each node, some additional external statistical analyses were considered in order to select the number of node correctly. The neural networks were implemented using Neural Network Toolbox for Matlab 6.5.

SI 5. Supplementary content to Section 2.6: Applicability domain

Applicability Domain

The leverage values are the representation of distance from the center of descriptor matrix which is derived based on the descriptors of training set compounds, as follows:

$$h_i = \mathrm{X_i^T}\left(\mathrm{X_i^T X}\right)^{-1} X_i \qquad (Eq.S10)$$

where $\mathrm{X_i}$ is the vector of descriptors for a compounds and X is the distance from the center of descriptor matrix derived from the training set. The warning leverage which can be an implantation of structural diversity is defining as follows:

$$h_i = 3k'/n \qquad (Eq.S11)$$

where n is the number of training set compounds, k' is the number of descriptors included in model plus one. In addition to the indication of higher leverages, *Williams plot* can show the compounds which are outliers based on their obtained retention times where compounds with cross-validated standardized residuals greater than three standard deviation units are being considered as outliers and should not be employed in models constructions. *Williams plot* is an implication of chemical diversity. There is another Euclidean based applicability domain which can be used to detect the outliers[22] only based on their structural dissimilarity. *Euclidean based applicability domain* helps to ensure that the compounds of the external set are representative of the training set compounds. This method is based on distance scores calculated by the Euclidean distance norms. Firstly, the normalized mean distance scores for training set compounds are calculated (these values ranges from 0 to 1 where 0.0 is least diverse, and 1.0 is the most diverse training set compound). Then, the normalized mean distance score for the test set compounds is calculated, and those test compounds obtained the scores outside of 0.0 to

1.0 ranges are defined to be outliers and cannot be predicted by the derived models. The final derived models where checked based on the two above models to verify their prediction abilities.

SI 6. Supplementary content to Section 2.7: External validation

External Validation

After the development of the models, it is highly needed to apply methods for evaluating the external predictive ability of the models. There are several external validation methods which can be used to validate the results, as discussed above (CCC and R^2 values). However, there is an important work performed by Tropsha who discussed the importance of the model validation.[23] As discussed, refereeing to Q^2_{LOO} and R^2 values for presenting the predictive ability of a built model is not enough for all cases, and the predictive power of a model can be investigated only based on the prediction results of the test set compounds. Therefore, an accurate and valid model can be established only based on model validation procedure consisted of properties prediction of compounds which were not included in the model development. Tropsha suggested that to simulate the use of QSAR/QSPR models, there should be another set of compounds with known activities/properties that are not included in either training or test sets. Then, by the proposed models, the activities of the built models are being predicted. In general, the size of the external validation set should be about 15% - 20% of the entire dataset, and the remaining part of the dataset, which is called modeling set, can be split into training and test sets. Following these procedures, Golbraikh and Tropsha acceptable model criteria's can also be a sufficient tool[24] to verify the predictive ability of the developed models. They introduced four conditions for accepting a model, as follows:

I) Q^2_{LOO} value must be higher than 0.5

II) R^2 value must be higher than 0.6

III) $R_0^2 - R_0'^2/R^2 < 0.1$ and $0.85 < K' < 1.15$ or $R^2 - R_0^2/R^2 < 0.1$ or $0.85 < K < 1.15$

IV) $R_0^2 - R_0'^2 < 0.3$

where R is the correlation coefficient between the predicted and observed values; R_0^2 is the coefficients of determination (correlation of predicted versus observed values with an intercept of zero), and $R_0'^2$ is the correlation between observed versus predicted values for regressions through the origin; K is the slope and K′ is the slope of the regression lines through the origin.[24] In this work, the above workflows for extra validation process were used, and some compounds as suspect screening were used as evaluation set for being predicted by the developed models. However, to be sure that the evaluation sets for both negative and positive ionizations are within the model applicability domain, Euclidean based applicability domain and OTrAMS were employed.

SI 7. Supplementary content to Section 3: Results and discussion

SI 7.1. Chromatographic system for the negative ionization analysis

PCA-stepwise-multiple linear regressions

After classification of the data set by PCA and kNN techniques into training and test set, the stepwise method was used to select the most respective molecular descriptors to understand the correlation of molecular structures with the retention times. Based on the stepwise method, the seven most relevant descriptors were selected and then the linear regression model was built. The linear model was developed based on the division of data set by PCA, as follows:

$$Rt = -0.4879\ (\pm 0.6701) - 0.5351(\pm 0.1141)\ nR06 + 0.9952(\pm 0.2119)\ ICR + 0.8935(\pm 0.2514)\ ATS3p - 0.6955(\pm 0.1018)\ EEig13d + 0.9912(\pm 0.1543)\ R3e + 0.5276(\pm 0.0767)\ ALOGP + 0.7372(\pm 0.06057)\ Log\ D_{(pH\ at\ 6.20)} \quad (Eq.S12)$$

N_{train}=239, R^2_{train}=0.854, $RMSE_{train}$=1.053, R^2_{adj}=0.850, F_{train}=195.523, Q^2_{LOO}=0.844, Q^2_{LGO}=0.777, Q^2_{BOOT}=0.842, N_{test}=59, R^2_{test}=0.782, $RMSE_{test}$=1.367, F_{test}=28.30, rm^2_{test} =0.724, CCC_{test}=0.8791, CCC_{train}=0.9216

where N is the number of compounds, R^2 is the squared correlation coefficient, R^2_{adj} is the adjusted R^2, $Q^2_{LOO,}$ Q^2_{BOOT} and Q^2_{LGO} are the squared cross-validation coefficients for leave one out, bootstrapping and leave group out, respectively. RMSE is the root mean square error and F is the Fisher F statistic. As it can be seen, the obtained model shows acceptable statistical parameters with higher square correlation coefficient (R^2), Fisher F statistic (F) and concordance correlation coefficient for both sets with lower RMSE

values. The predicted retention time values for the whole range of the compounds in training and test sets using the equation 9 have been plotted against the observed retention time values in **Figure S2**, and listed in **Table S1**. The corresponding VIF values and inter-correlation values of the selected seven descriptors are shown in **Table S2**. As can be seen from this Table, all variables have VIF values less than 5, indicating that the obtained model has appropriate selected variables. Also low R^2 and Q^2 values were obtained by Y-randomization test (**Table S3**). The robustness of the proposed model and its predictive ability was guaranteed by the high Q^2_{BOOT} based on bootstrapping repeated 5000 times. Applicability domain was used and outliers were detected and removed. Two outliers were found that possessed residuals more than ±3δ (**Figure S3**). These two compounds belonged to the test set, and were not included in the model development; therefore, their omission just benefits the outcome of the correlation in the test set (R^2 from 0.782 to 0.818). Before interpreting the descriptors based on PCA-SW-MLR, Genetic algorithms technique is also used to compare the results of the two methods.

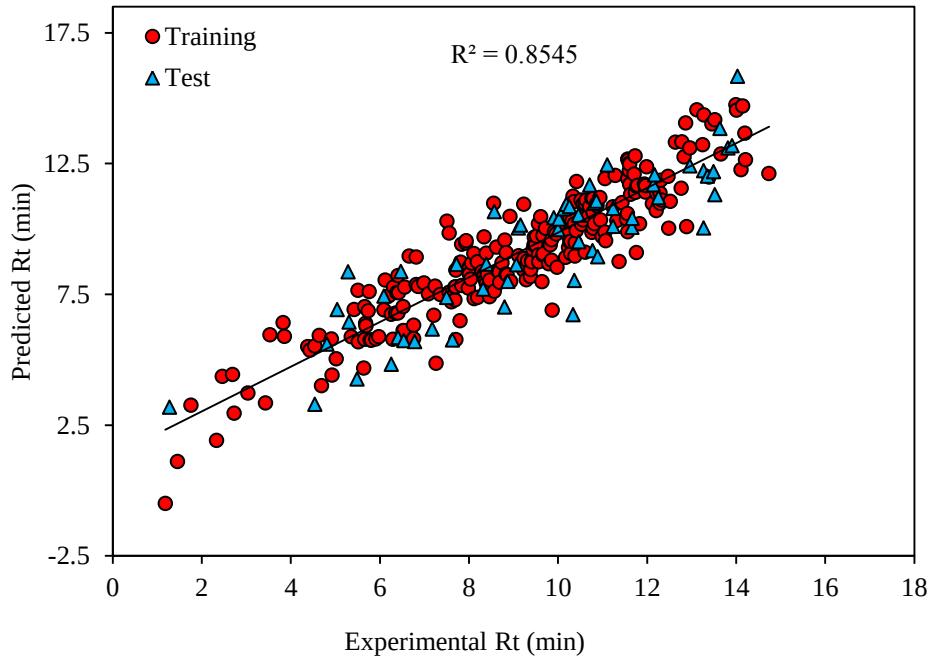


Figure S2. The plot of predicted retention time against the observed retention time values based on PCA-SW-MLR

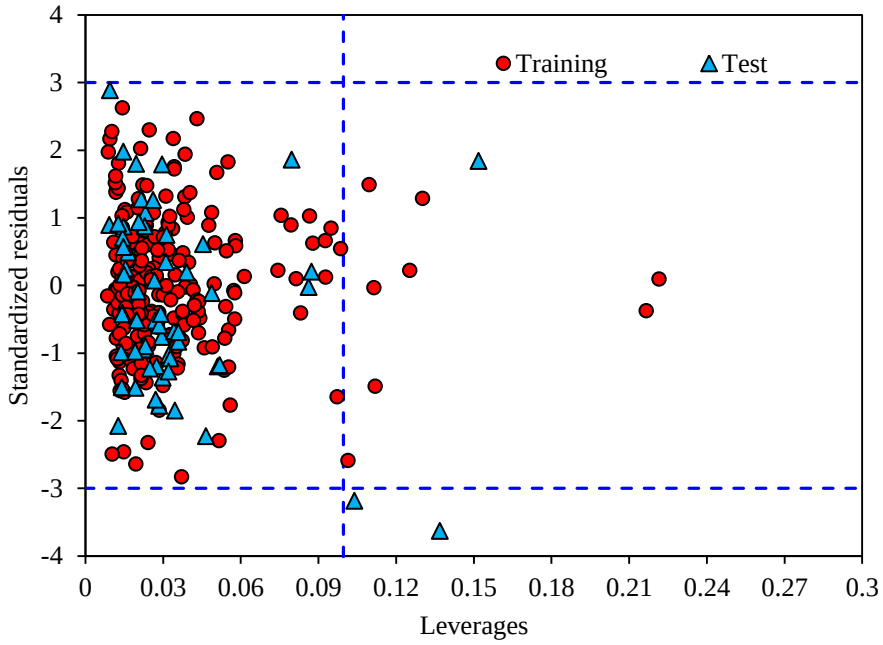


Figure S3. Williams plot of PCA-SW-MLR model (equation S12): h* warning leverage value is 0.09985.

PCA-genetic algorithms-multiple linear regressions

After classification of the data set by the same procedure done in PCA-SW-MLR, genetic algorithms were used to select the most relevant descriptors. For selecting the best subset of descriptors, GAs were applied for several times and the model presented the higher statistical parameters was chosen. The results of combinations of different couples of descriptors selected by GAs were listed in **Table S4**.

Rt = 1.789(±0.5342) + 0.6561(±0.06779) $logD_{(pH\ at\ 6.20)}$ - 0.5386 (±0.08216) BLTA96 + 0.3083(±0.06447) AlogP + 0.1038(±0.1205) nROH + 0.5174 (±0.376) HATS6m + 1.591(±0.29003) R2e - 0.2762 (±0.2209) Mor25e $(Eq.S13)$

R^2_{train}=0.812, $RMSE_{train}$=1.201, R^2_{adj}=0.806, F_{train}=143.94, Q^2_{LOO}=0.789, Q^2_{LGO}=0.700, Q^2_{BOOT}=0.788, R^2_{test}=0.786, $RMSE_{test}$=1.379, F_{test}=28.08, rm^2_{test} =0.721, CCC_{test}=0.8775, CCC_{train}=0.8960

The obtained statistical parameters (high squared correlation coefficient, CCC, Q^2_{BOOT} and Q^2_{LOO}) show that genetic algorithms technique is better than stepwise method for selecting the descriptors as model variables. To find out that the selected descriptors statistically meaningful, the Y-randomization test (**Table S5**) was used. In this method, properties for group of compounds were shuffled, and then, a new model was built. The new QSPR models as outcome of this method should present the low R^2 and Q^2_{LOO} values in order to be confident that the models are directly in relation with the selected variables. The predicted retention time values for the whole compounds in training and test sets using the equation 10 plotted against the observed retention time values in **Figure S4**, and listed in **Table S1**. The corresponding VIF values and inter-correlation values of the selected seven descriptors are listed in **Table S6**. As can be seen from this Table, all variables have VIF values less than 5, indicating that the obtained model has acceptable descriptors. The applicability domain of the generated model was also studied showing no outliers that possessed the residuals more than ±3δ (**Figure S5**). The PCA-GA-MLR model (Eq 10) was obtained after the removal of semduramicin and alitame, and the second built model did not show any outliers for training set. Some other compounds were located outside the warning leverage value, however they did not show high residuals (more than ±3δ), and therefore they were not treated as outliers. To understand the reason why these two compounds are outliers, the molecular descriptors which were selected by GAs can be used as input for Euclidean based applicability domain (**Figure S6**) so as to explain the diversity of compounds based on the selected descriptors. As it can be seen, the origin of outliers are not derived from the structural diversity, since they are within the capability of the model to be predicted, however the observed retention time don't match to the given structure. *Therefore, the PCA-GA-MLR model can be accepted as an initial model for predicting purposes*. This workflow can help to understand if the screened unknown and suspect compounds to be studied further are within the capability of the models or not, before predicting their values.

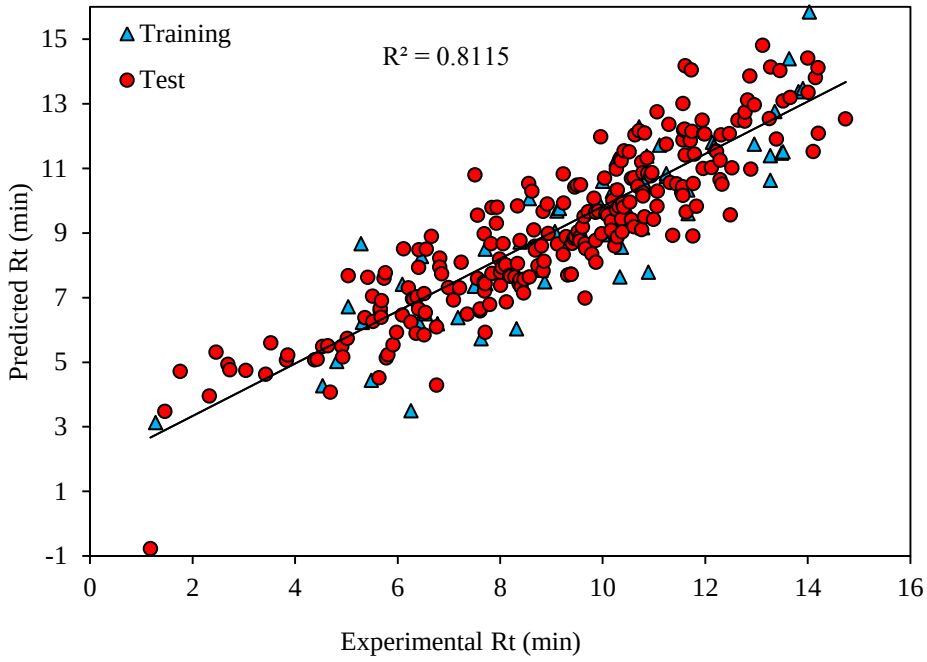


Figure S4. The plot of predicted retention time against the observed retention time values based on PCA-GA-MLR

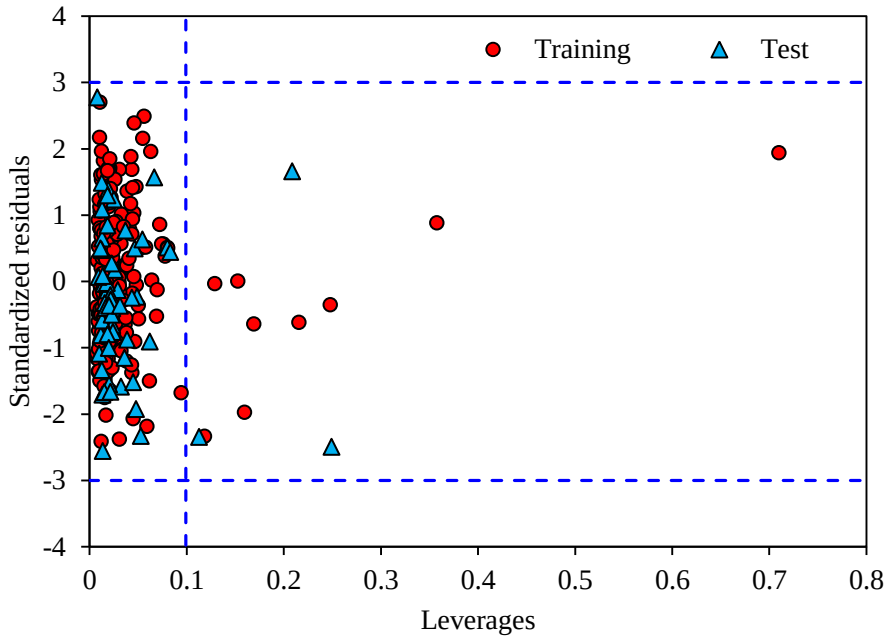


Figure S5. Williams plot of PCA-GA-MLR model: h* warning leverage value is 0.09917.

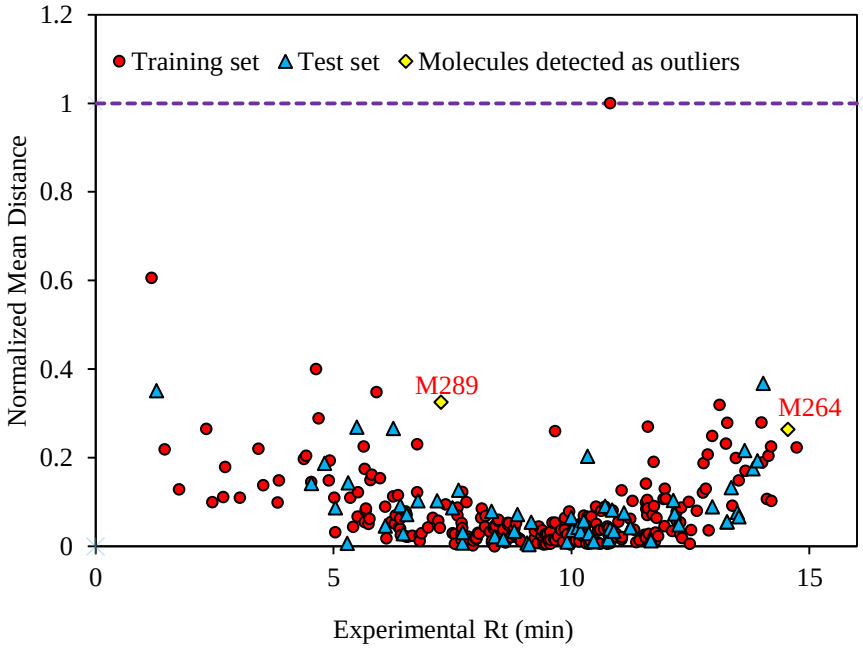


Figure S6. Euclidean based applicability domain of the compounds for PCA-GA-MLR

PCA-stepwise-support vector machine

After the development of successful linear models based on both stepwise and genetic algorithms techniques, support vector machines were used as non-linear modeling technique on the same subsets of descriptors. As explained before, SVM regression depends on the combination of different factors such as kernel function type, capacity parameter *C*, ε of ε-insensitive loss function, and its corresponding parameters.[25] For generating the SVM model, firstly, the Kernel function type should be declared in which determines the sample distribution in space. In this work, the radial basis function (RBF) was used due to its good general performance.[26] Gamma (γ) is in close relation with SVM performance (its training time) and controls the generalization ability of SVM. Generally, to get the optimum value for γ, it is being measured from 0.1 to 5 with incremental steps of 0.1. To get better insight about the optimized values, the root mean square errors (RMSE) of cross-validation were obtained in each step. **Figure S7** represents the plot of γ versus the RMSE on the leave-one-out cross-validation results. Here the optimal value of 2.7 has been obtained for γ.

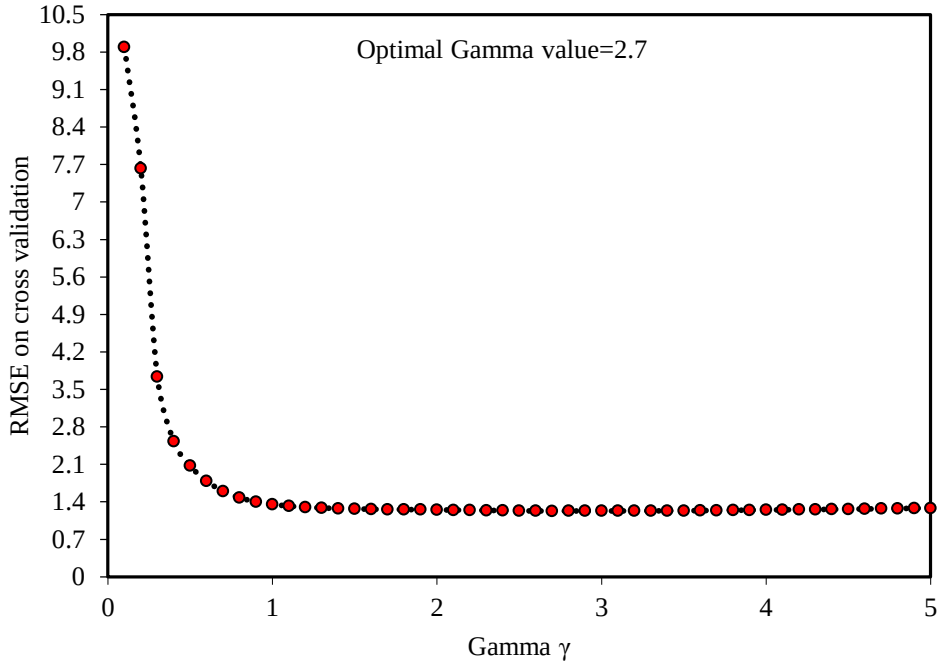


Figure S7. The gamma(γ) vs. RMSE for the training set based on PCA-SW-SVM.

Parameter ε-insensitive prevents the entire training set meeting boundary conditions, and so allows for the possibility of sparsity in the dual formulation's solution. The optimal value for ε depends on the type of noise present in the data, which is usually unknown. ε-insensitive has an effect over smoothness of the response of SVM, and also influence the number of support vectors. An increase in ε-insensitive value reflects the reduction in requirements for the desired accuracy approximation. Therefore, if ε-insensitive is zero, there is an over-fitting issue, and if it presents larger values than the range of target values, the obtained results are not appropriate. The RMSEs of cross-validation for different ε values from 0.01 to 0.1 with incremental steps of 0.01 are shown by **Figure S8**. The optimal value for ε-insensitive is 0.01.

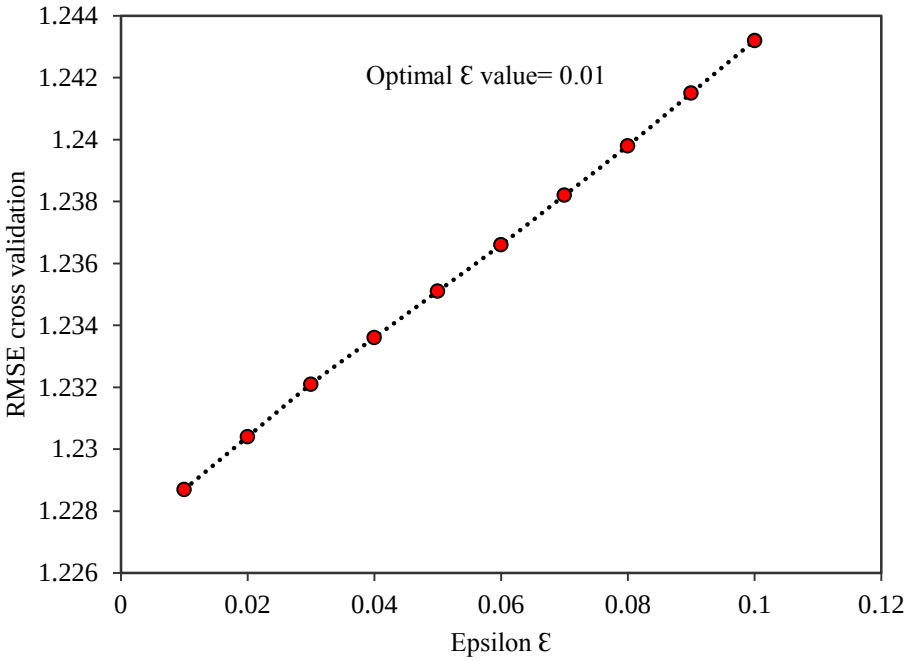


Figure S8. The epsilon (ε) vs. RMSE for the training set based on PCA-SW-SVM.

The final parameter which should be optimized was C that is a regularization parameter that controlled the trade-off between maximizing the margin and minimizing the training error. The small values for C parameter would increase the number of training errors, and a large value would cause hard-margin SVM behavior. The capacity parameter C was checked from 1 to 50 with incremental steps of 1 and is shown in **Figure S9**. The optimal value for capacity parameter is 50.

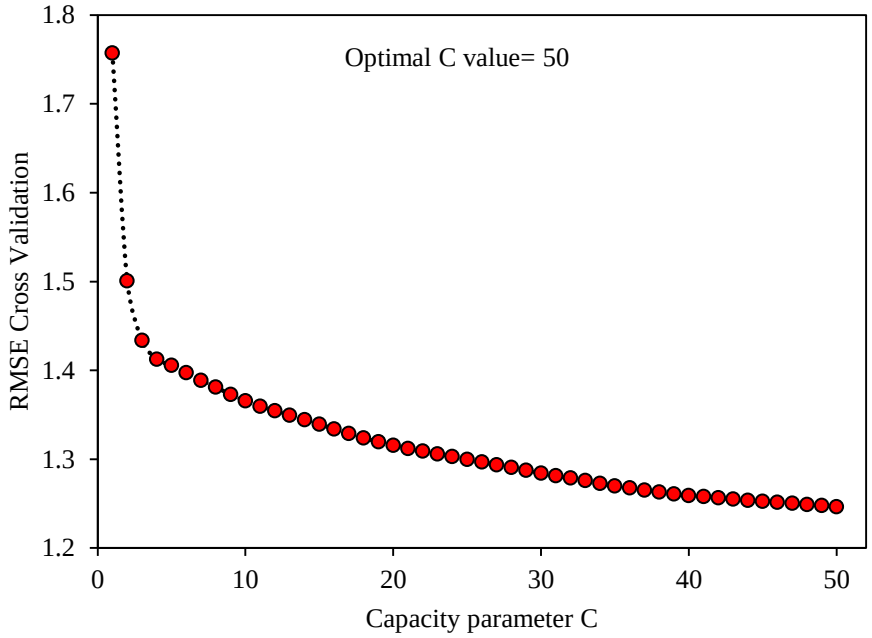


Figure S9. The capacity parameter (C) *vs* RMSE for the training set based on PCA-SW-SVM.

The parameters of SVM model were optimized as C=50, ε=0.01, γ=2.7. The predicted values for retention time by SVM were given in **Table S1**. Also, the predicted *versus* experimental retention time values are shown in **Figure S10** for both the training and the test set based on SVM model. The statistical parameters for PCA-SW-SVM model showed RMSE values of 0.486 for the training set, 1.25 for the test set, and squared correlation coefficients (R^2) of 0.970 and 0.818 for training and test set, respectively. **Table 1** presents the statistical parameters of the results obtained from the studied models for the same set of compounds. For obtaining better results, the above workflow was performed for compounds of training and test sets, selected by k-nearest neighborhood clustering technique.

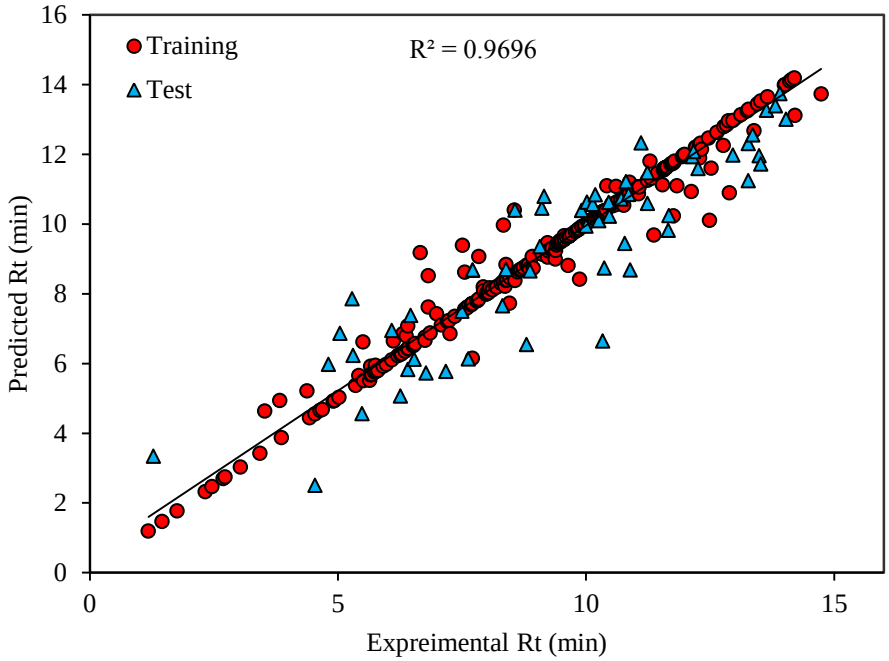


Figure S10. The plot of predicted retention time against the observed retention time values based on PCA-SW-SVM

PCA-genetic algorithm-support vector machine

The same procedure employed in PCA-SW-SVM model was performed in this part; however the non-linear model was built based on the selected descriptors using genetic algorithms as a selection tool. The test set compounds were marked in **Table S1** which were the same used in generation of PCA-GA-MLR model. The parameters of SVM model were optimized as C=50, ε=0.01, γ=1.9. The results of each optimization were shown in **Figure S11 (A-C).** The predicted values for retention time by PCA-GA-SVM model were given in **Table S1**, and then plotted versus the observed retention time in **FigureS12**. The statistical results of PCA-GA-SVM were listed in **Table 1**.

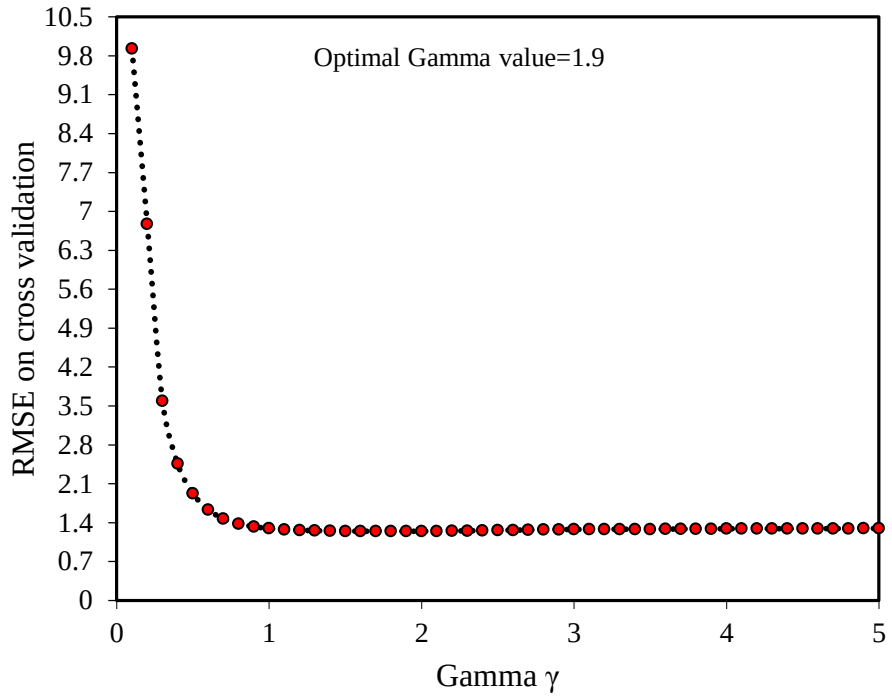


Figure S11 (A)**.** PCA-GA-SVM optimized parameters for the gamma (γ) vs. RMSE

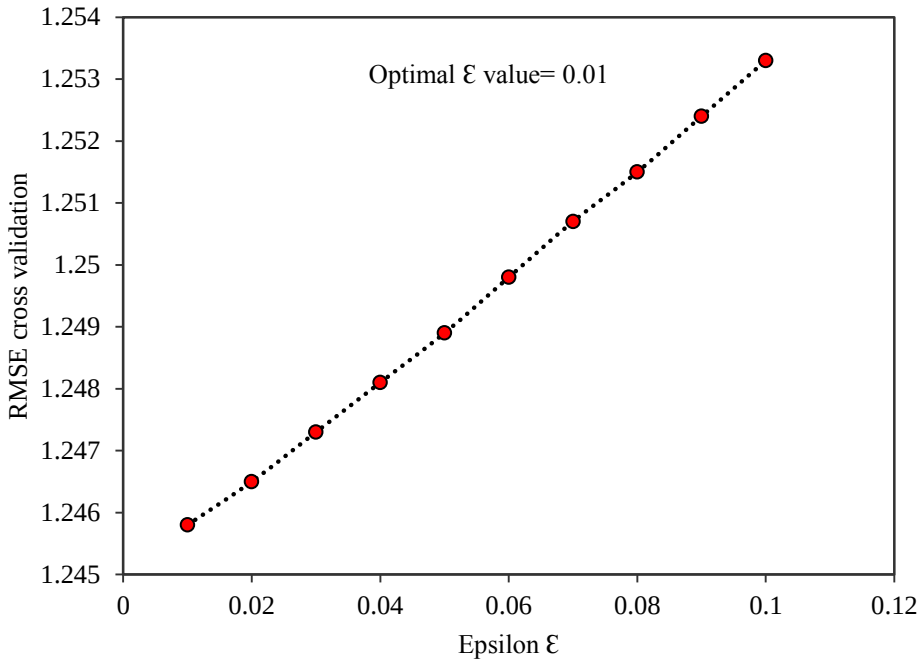


Figure S11 (B)**.** PCA-GA-SVM optimized parameters for the epsilon (ε) vs. RMSE

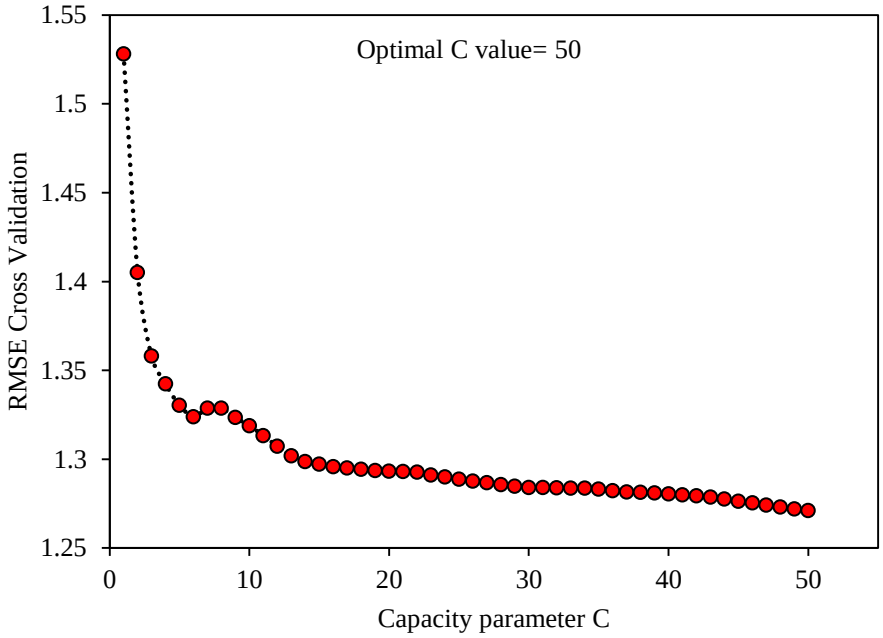


Figure S11(C)**.** PCA-GA-SVM optimized parameters for the capacity (C) vs. RMSE

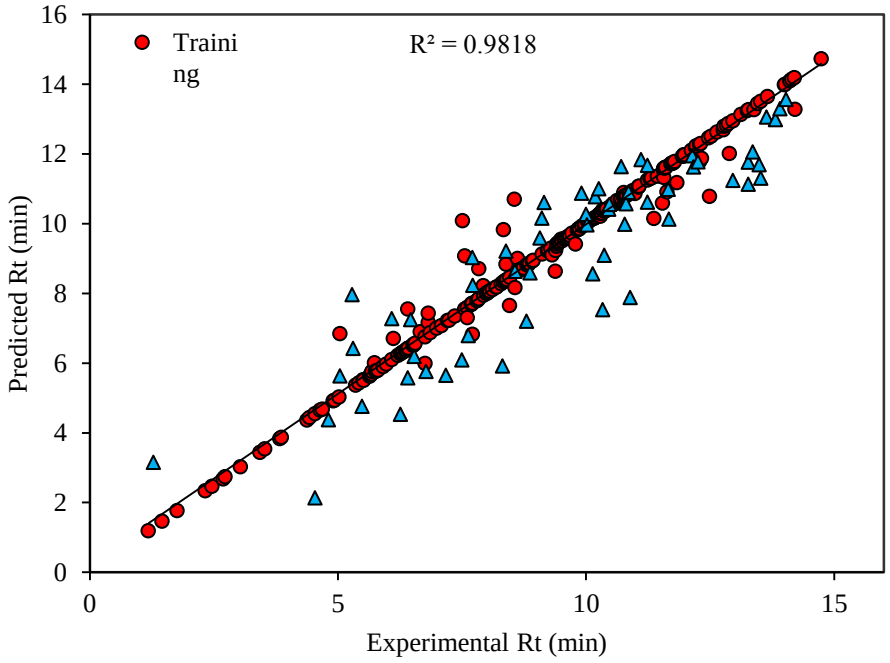


Figure S12. The plot of predicted retention time against the observed retention time values based on PCA-GA-SVM.

kNN-SW-MLR and kNN-SW-SVM

The same procedures were used for developing the linear and non-linear models; however, the data set was spilt based on the results of the dendrograms. The test compounds were marked in **Table S1** and were shown in **Figure S1 (C)**. Since the interpretations of results were explained above, here we are just presenting the obtained results for kNN-SW-MLR model. The linear model was calculated as follows:

$$Rt = -0.5622\ (\pm 0.6977) + 0.9148\ (\pm 0.2304)\ ATS3p + 1.657(\pm 0.3289)\ GATS2m - 1.006(\pm 0.1392)\ EEig14r + 1.601\ (\pm 0.2167)\ R3u + 0.4980(\pm 0.07977)\ AlogP - 0.7737(\pm 0.1948)\ B02[C\text{-}S] + 0.7197(\pm 0.0623)\ \log D_{(pH\ at\ 6.20)} \qquad (Eq.S14)$$

N_{train}=239, R^2_{train}=0.842, $RMSE_{train}$=1.107, R^2_{adj}=0.837, F_{train}=176.95, Q^2_{LOO}=0.829, Q^2_{LGO}=0.752, Q^2_{BOOT}=0.827, N_{test}=59, R^2_{test}=0.822, $RMSE_{test}$=1.209, F_{test}=29.41, rm^2_{test} =0.770, CCC_{test}=0.8941, CCC_{train}=0.9140

The Y-randomization test was used, and the results were indicated that developed model is acceptable (**Table S7**). *Williams plot* was also calculated to detect the possible outliers; however outliers were not detected for the training set (**Figure S13**). Only one molecule which belonged to the test set was detected as outlier, in which its omission will not benefit the model, since it was not included in model construction. The predicted retention time values using the equation S14 plotted against the observed retention time values in **Figure S14**, and the results for the whole data set were listed in **Table S1**. The non-linear model was built based on the selected compounds as training set as same as kNN-SW-MLR, and the optimized parameters were obtained as C=7.0, ε=0.1, γ=2.1. The result of each optimization was shown in **Figure S15 (A-C)**. The predicted values for retention time by kNN-SW-SVM method were given in **Table S1**, and then plotted *versus* the observed retention time and shown in **Figure S16**. The statistical results of this model were listed in **Table 1**.

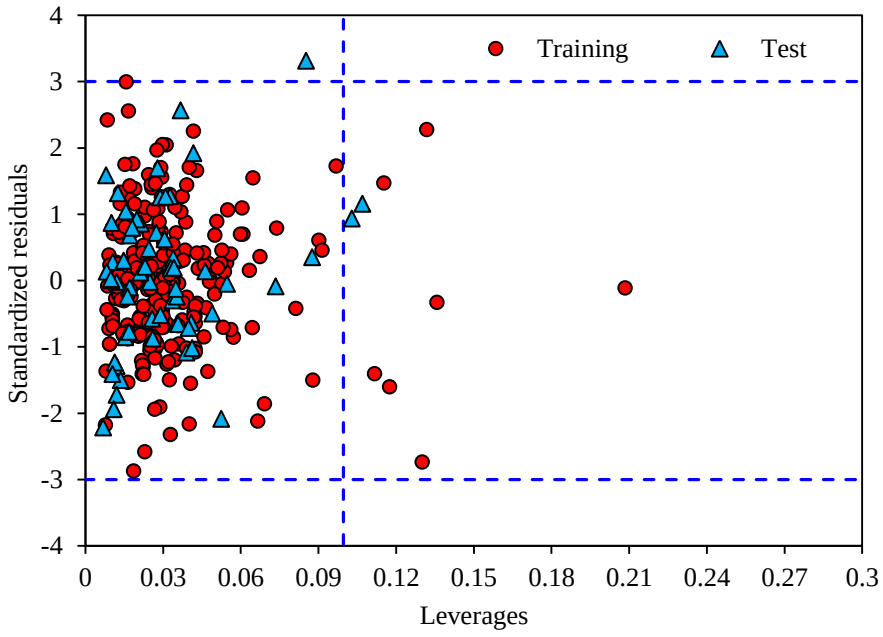


Figure S13. *Williams plot* of kNN-SW-MLR model: h* warning leverage value is 0.099585.

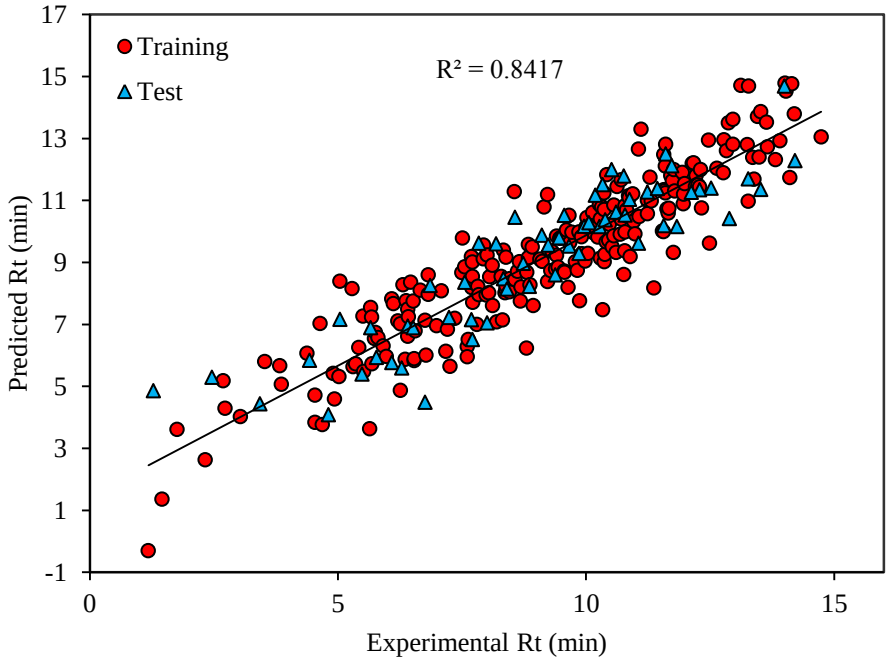


Figure S14. The plot of predicted retention time *versus* the observed retention time values based on kNN-SW-MLR.

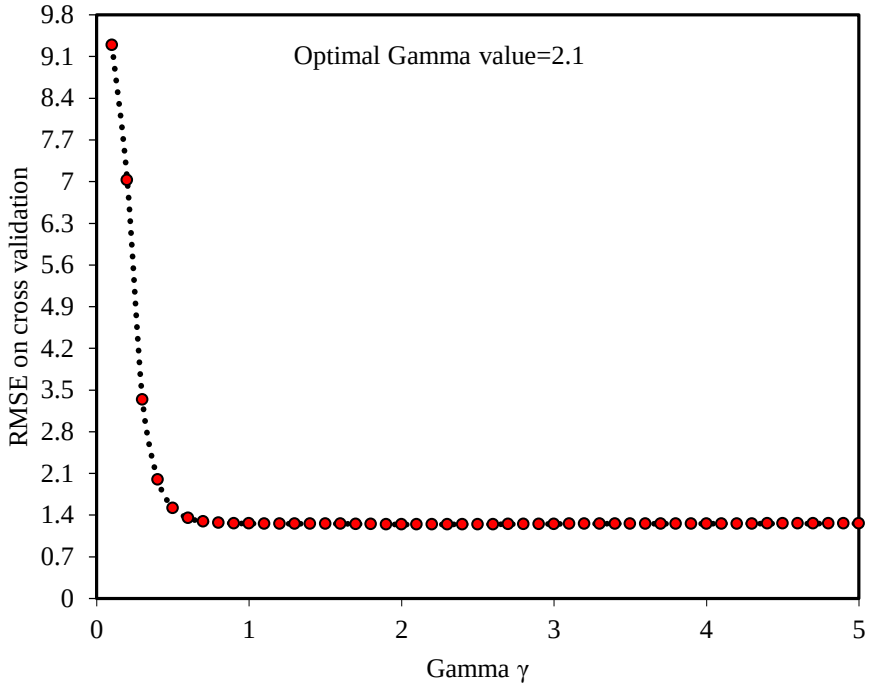


Figure S15 (A)**.** kNN-SW-SVM optimized parameters for the gamma (γ) *vs* RMSE.

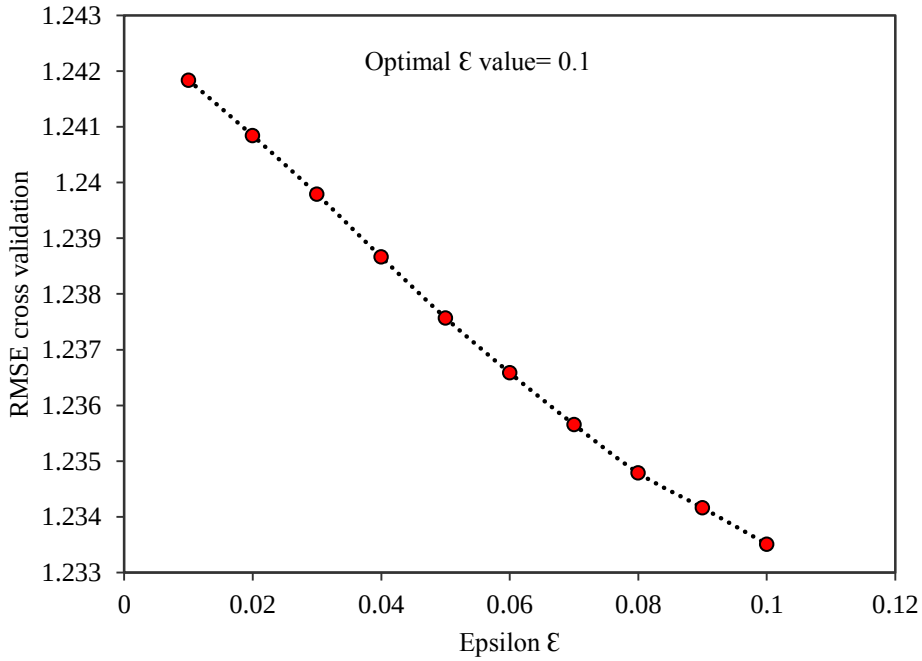


Figure S15 (B)**.** kNN-SW-SVM optimized parameters for the epsilon (ε) *vs* RMSE.

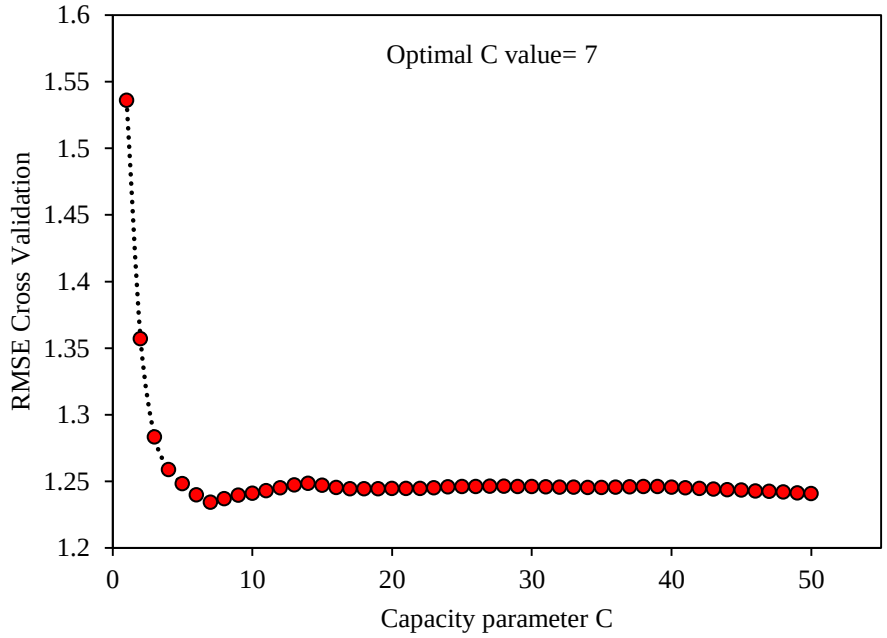


Figure S15 (C)**.** kNN-SW-SVM optimized parameters for the capacity (C) *vs* RMSE

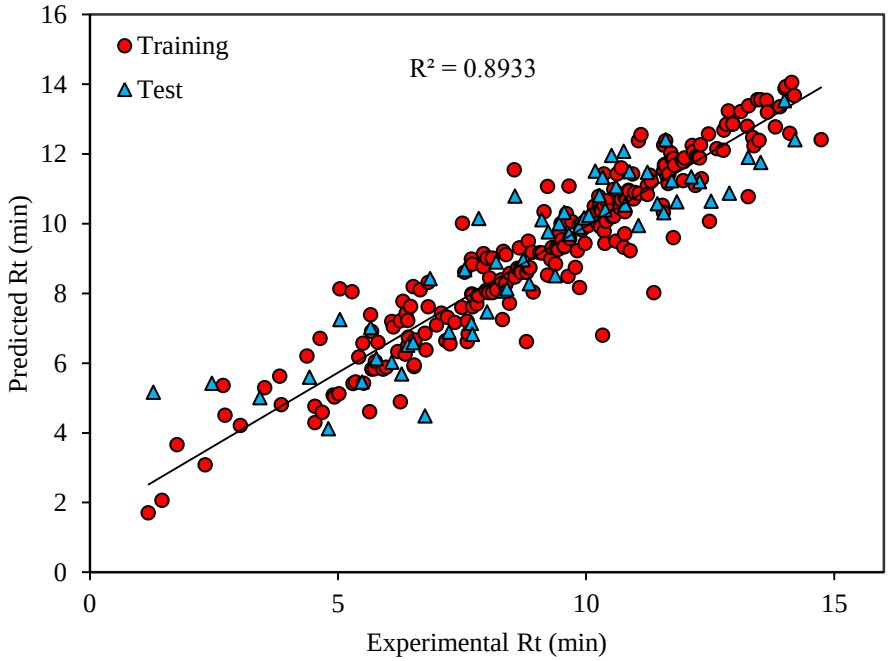


Figure S16. The plot of predicted retention time *versus* the observed retention time values based on kNN-SW-SVM.

kNN-GA-SVM

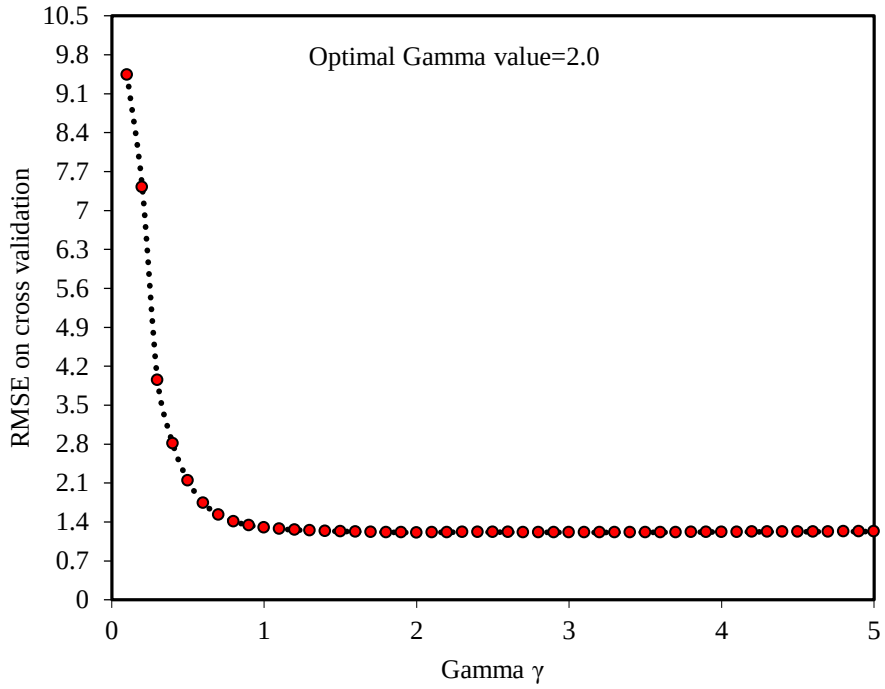


Figure S17 (A). kNN-GA-SVM optimized parameters for the gamma (γ) *vs* RMSE.

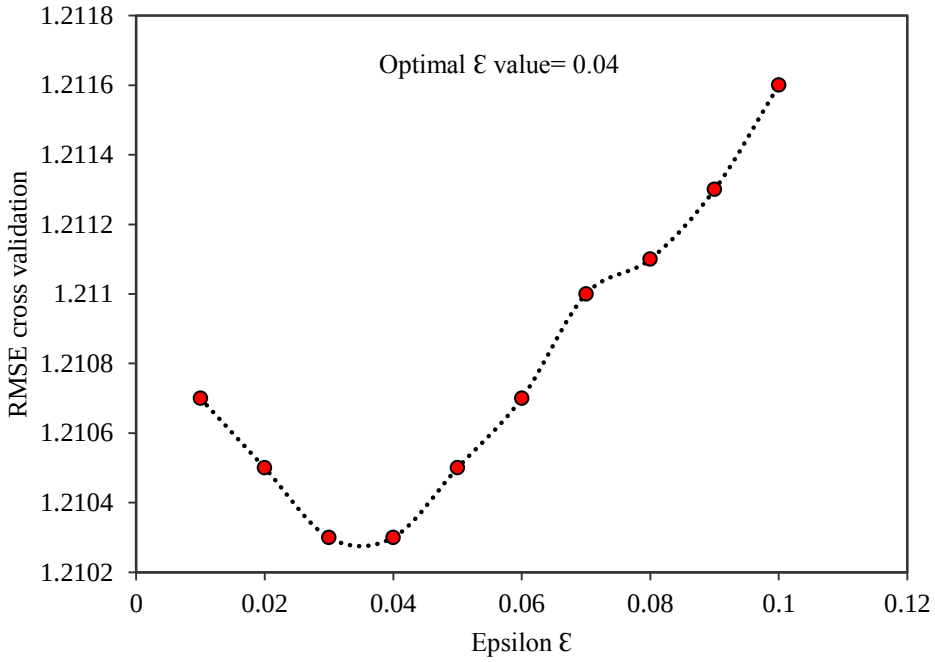


Figure S17 (B). kNN-GA-SVM optimized parameters for the epsilon (ε) *vs* RMSE.

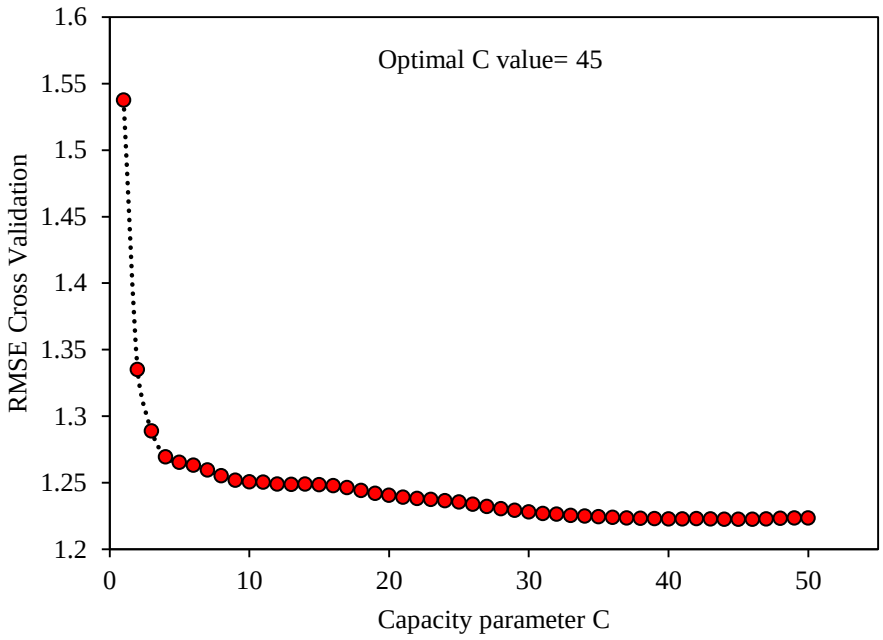


Figure S17 (C). kNN-GA-SVM optimized parameters for the capacity (C) *vs* RMSE

kNN-GA-ANN

Since the models based on kNN and genetic algorithm showed appropriate internal and external results, the non-linear model based on ANN was developed based on kNN-genetic algorithm technique. It is accepted that for the generation of ANN models employing variable selection is not necessary, but it can be useful to get better results. Therefore, we used genetic algorithm for descriptors subset selection in ANN. The common problem with ANN is to select the right node where the RMSE values are being considered for final model construction. Here, we reported and selected the ANN model based on the modified r^2 value, CCC value, and RMSE. Therefore, considering the over-fitting problem in higher nodes, the right nodes can be selected using their CCC values first that encodes the accuracy and precision, and then provided modified r^2 value for test set to select the nodes. Finally, for the couple of nodes with acceptable results for the test set, the one that is showing also less RMSE value for the training set can be selected as the final node for subsequent analysis. The results of this procedure are shown in **Table S10**. From this table, it can be seen that the model built based on node=7 shows the highest CCC value for both the test and the training set, and the modified r^2 value for test

set is the highest one among the other nodes. Considering the RMSE value between node 5 and 7, consequently the model based on 7 nodes is being selected. The MPD values for training set were calculated for the all nodes as follows:

$$MPD = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (Eq.S15)$$

where $y_i$is the observed retention time, and $\hat{y}_i$ the calculated retention time and N denotes the number of data points. The results are given in **Table S10.** This formula measures the accuracy of the generated models based on each node and the lower value indicate the good fitted point. The predicted values based on kNN-GA-ANN are listed in **Table S1** and their strength as prediction tool is compared with the other models in **Table 1**. The correlation plot of observed and predicted retention time is shown in **Figure S18**.

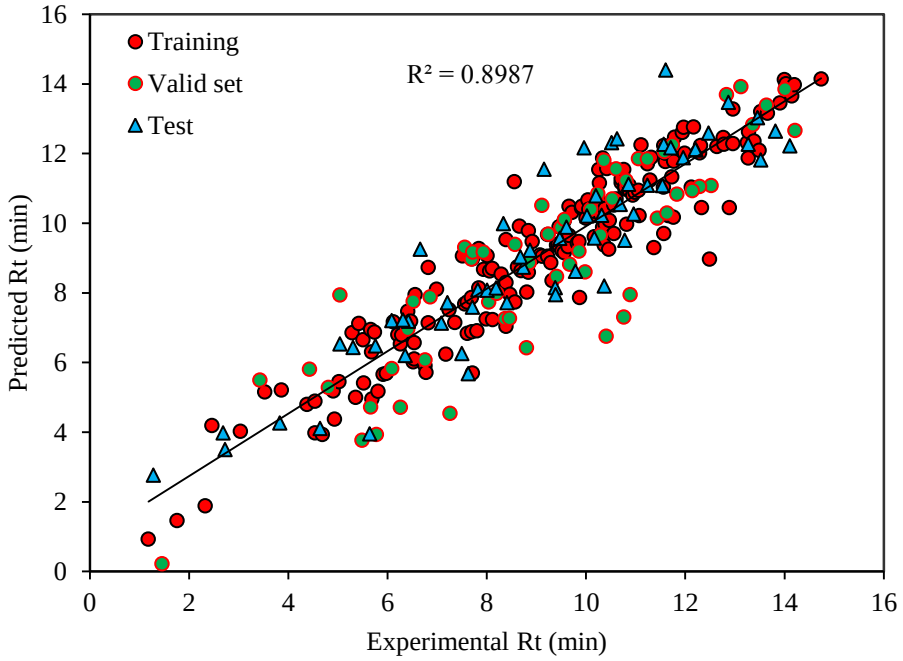


Figure S18. The plot of predicted retention time against the observed retention time values based on kNN-GA-ANN

SI 7.2. Chromatographic system for the positive ionization analysis

PCA-SW-MLR and PCA-GA-MLR

Since the workflow was the same as that employed in negative ionization, here the results of each step were discussed in less detail. The linear models based on stepwise variable selection tool and selected test set compounds by PCA and kNN are calculated initially before performing the genetic algorithms technique. The selected compounds as test set based on each splitting techniques were shown in **Table S1** (positive ionization). The model based on the PCA-SW-MLR is as follows:

Rt= 2.021 (±1.351) + 1.899(±1.897) Mv + 0.1021(±0.0291) RBN + 0.8486(±0.1384) CIC1 -0.3978(±0.05838) C-025 + 0.0513(±0.01264) MlogP2 + 1.685(±0.2639) B06[C-C] + 1.097(±0.05665) $logD_{(pH=3.6)}$ $(Eq.S16)$

N_{train}=422, R^2_{train}=0.846, $RMSE_{train}$=1.061, R^2_{adj}=0.843, F_{train}=324.49 Q^2_{LOO}=0.840, Q^2_{LGO}=0.478, Q^2_{BOOT}=0.838, N_{test}=105, R^2_{test}=0.843, $RMSE_{test}$=1.127, F_{test}=78.49, rm^2_{test} =0.765, CCC_{test}=0.9127, CCC_{train}=0.9165

The Y-randomization test was calculated, and the results were indicated that developed model is acceptable (**Table S13**). *William plot* detected 3 outliers (1 for training and 2 for test set) in model. After removal of the detected outlier from the training set, the final predictive model obtained (**Figure S19**). VIF values for each selected descriptor along with correlation values between pair descriptors are listed in **Table S14**. The predicted retention time values using the equation S16 plotted versus the observed retention time values in **Figure S20**. The linear model based on genetic algorithms is also developed to compare the results. The model based on the PCA-GA-MLR is as follows:

Rt= 3.559(±0.3267) +0.9348(±0.06201) $logD_{(pH=3.6)}$ -0.2956 (±0.0704) BLTA96 +0.1394 (±0.02849) RBN +0.00408(±0.00926) AlogP2 -0.2621(±0.0686) nHDon +0.5871 (±0.1086) CIC1 +1.282(±0.2610) B06[C-C] $(Eq.S17)$

N_{train}=421, R^2_{train}=0.849, $RMSE_{train}$=1.048, R^2_{adj}=0.846, F_{train}=331.19 Q^2_{LOO}=0.842, Q^2_{LGO}=0.731, Q^2_{BOOT}=0.841, N_{test}=104, R^2_{test}=0.816, $RMSE_{test}$=1.154, F_{test}=59.00, rm^2_{test} =0.814, CCC_{test}=0.8966, CCC_{train}=0.9182

The Y-randomization test and VIF values were given in **Table S15 and S16**, respectively. *William plot* detected 2 outliers for (1 for training and 1 for test set) in model. After removal of the detected outlier from the training set, the final predictive model obtained (**Figure S21**). The predicted retention time values *versus* the observed retention time values for PCA-GA-MLR model are presented in **Figure S22**.

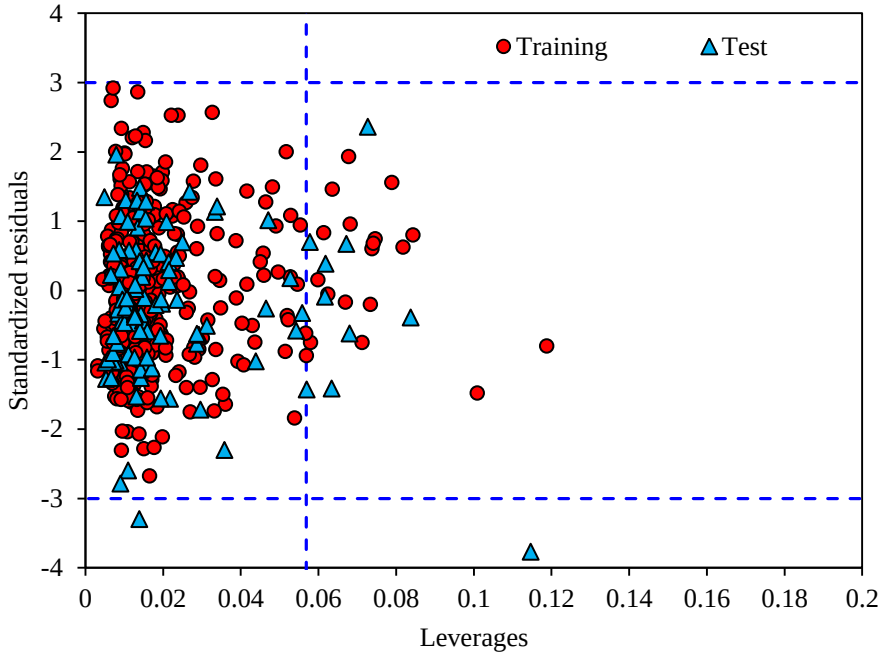


Figure S19. Williams plot of PCA-SW-MLR model: h* warning leverage value is 0.056872.

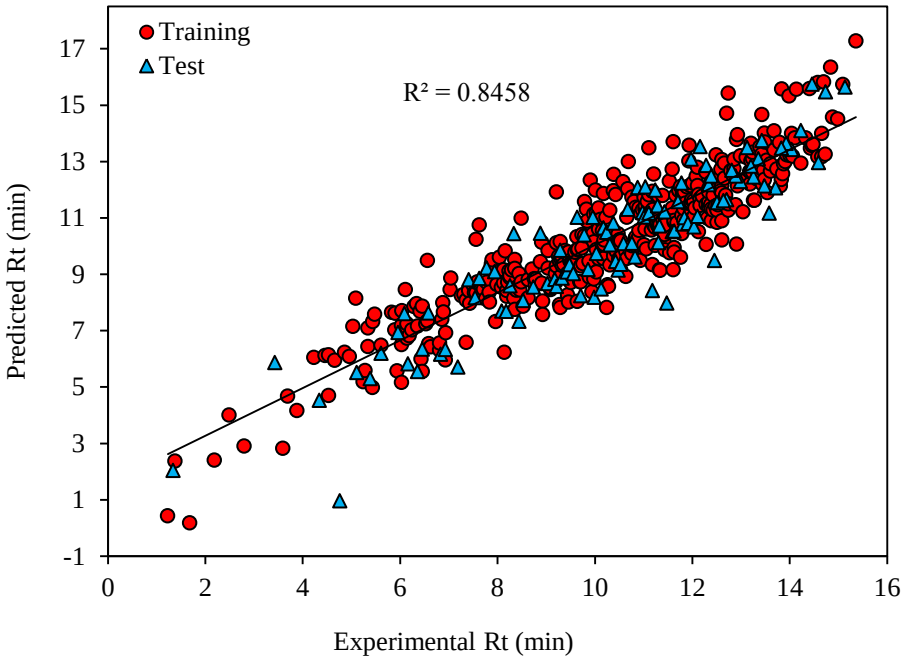


Figure S20. The plot of predicted retention time against the observed retention time values based on PCA-SW-MLR (positive ionization)

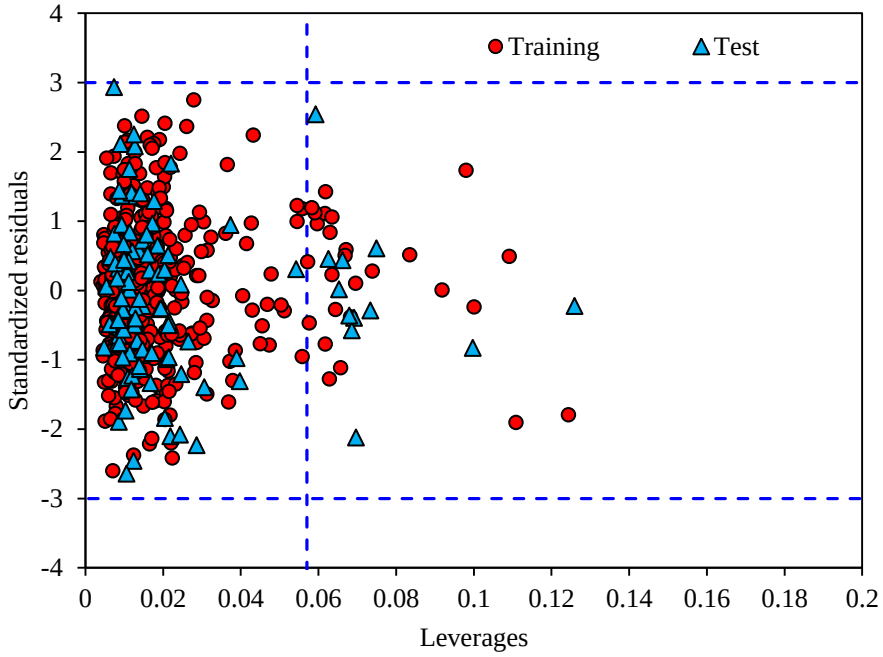


Figure S21. Williams plot of PCA-GA-MLR model: h* warning leverage value is 0.057007.

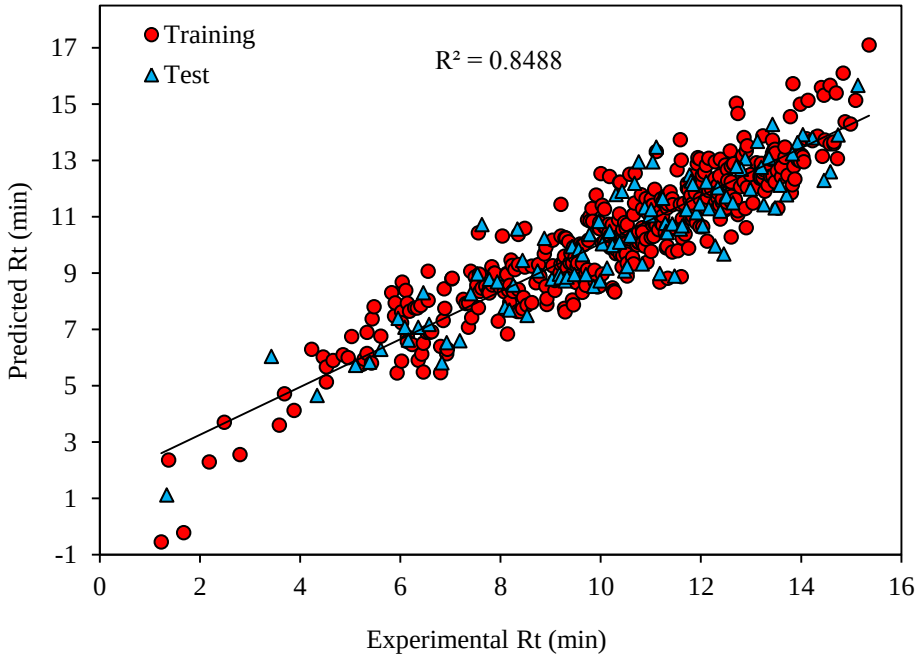


Figure S22. The plot of predicted retention time against the observed retention time values based on PCA-GA-MLR (positive ionization)

The same procedures were done for dataset split by kNN technique. The results for kNN-SW-MLR were obtained as follows:

Rt= 2.159(±1.367) +2.126(±1.9101) Mv +0.1228 (±0.0277) RBN +0.7832(±0.1381) CIC1 -0.409 (±0.0575) C-025 +0.0500(±0.0128) MlogP2 +1.501(±0.23142) B06[C-C] +1.0854 (±0.0551) $logD_{(pH=3.6)}$ $(Eq.S18)$

N_{train}=422, R^2_{train}=0.847, $RMSE_{train}$=1.051, R^2_{adj}=0.844, F_{train}=327.20 Q^2_{LOO}=0.841, Q^2_{LGO}=0.744, Q^2_{BOOT}=0.840, N_{test}=105, R^2_{test}=0.826, $RMSE_{test}$=1.162, F_{test}=67.03, rm^2_{test} =0.809, CCC_{test}=0.9050, CCC_{train}=0.9171

VIF values for each selected descriptor along with correlation values between pair descriptors were listed in **Table S17**. *William plot* detected two outliers for final kNN-SW-MLR model (2 compounds for test set) (**Figure S23**). The predicted retention time versus the observed retention time values based on kNN-SW-MLR were shown in **Figure S24.**

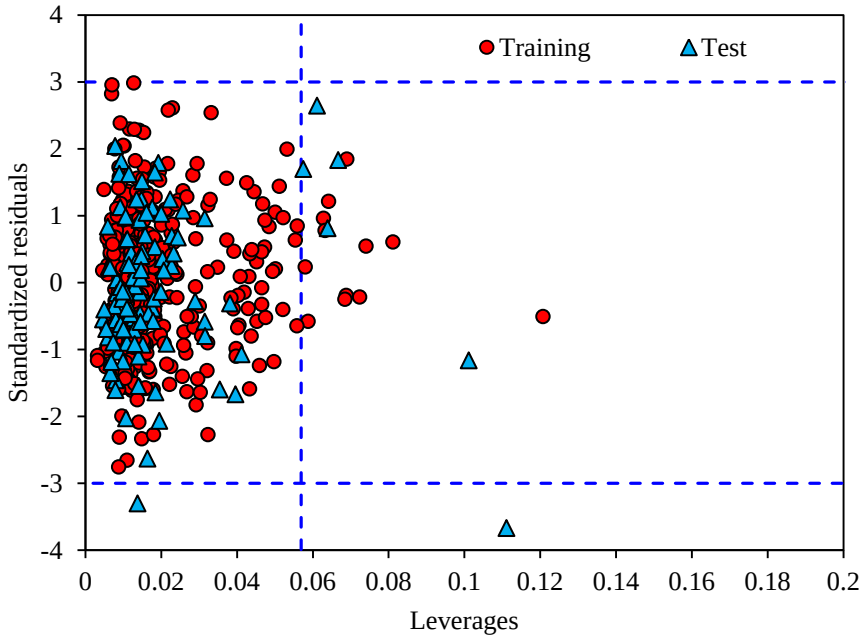


Figure S23. *Williams plot* of kNN-SW-MLR model: h* warning leverage value is 0.05687, namely.

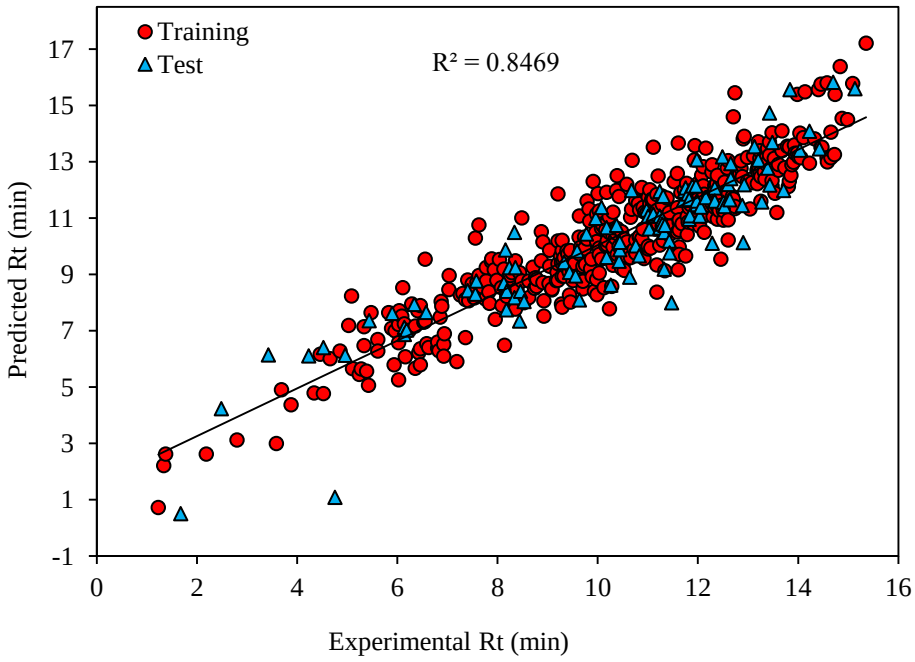
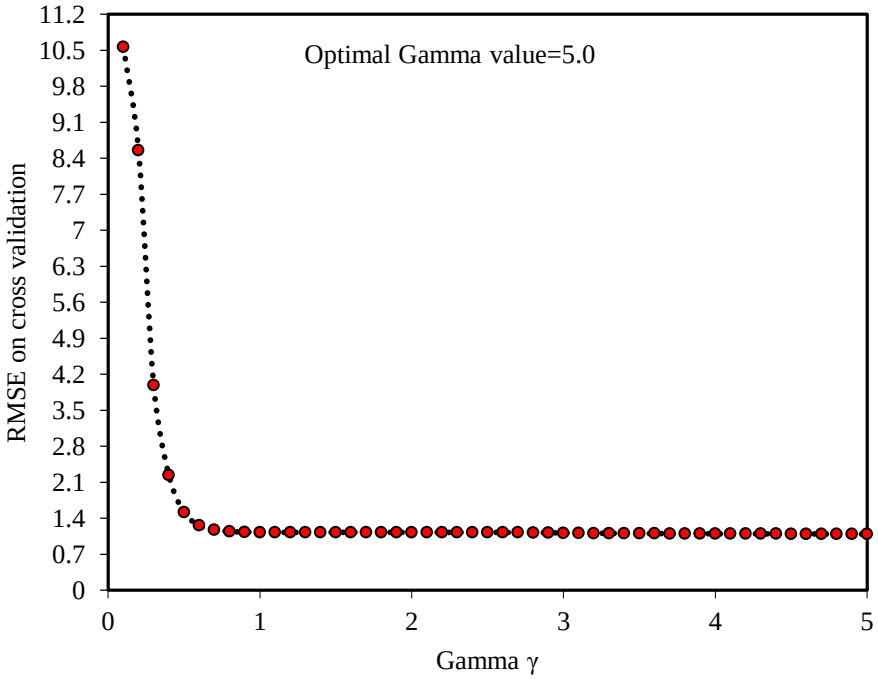


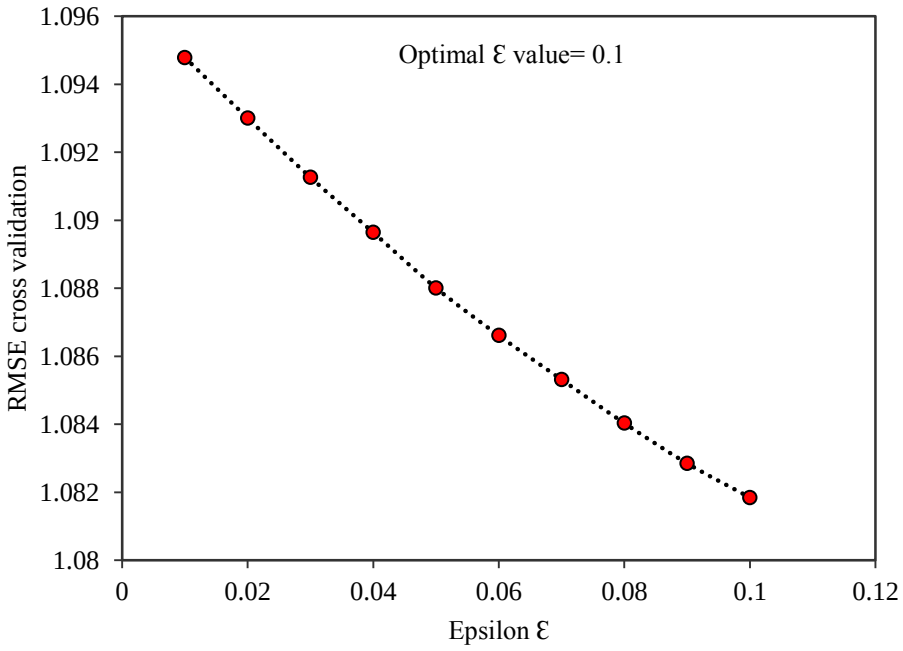
Figure S24. The plot of predicted retention time against the observed retention time values based on kNN-SW-MLR model.

SVM models

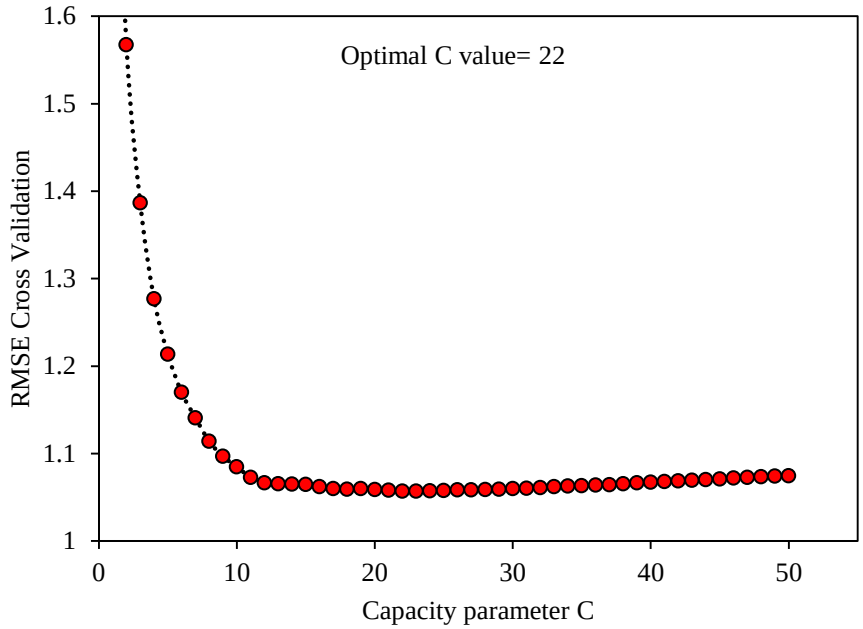
The used methodology for developing the SVM discussed in negative ionization was employed here as well. The results of optimization of parameters for each SVM models based on stepwise and genetic algorithms with different splitting technique were listed in **Table S20** and were shown in **Figure S25**. For each non-linear model, the results were listed in **Table S1** (positive ionization). The results of each different model were compared to linear models and presented in **Table 4**. As it can be seen from **Table 4**, the best non-linear model was obtained based on the kNN-GA-SVM. The plots of predicted retention time versus the observed retention time values based on SVM methods are shown in **Figure S26**.


Optimal Gamma value=5.0
RMSE on cross validation
Gamma γ
11.2
10.5
9.8
9.1
8.4
7.7
7
6.3
5.6
4.9
4.2
3.5
2.8
2.1
1.4
0.7
0
0
1
2
3
4
5

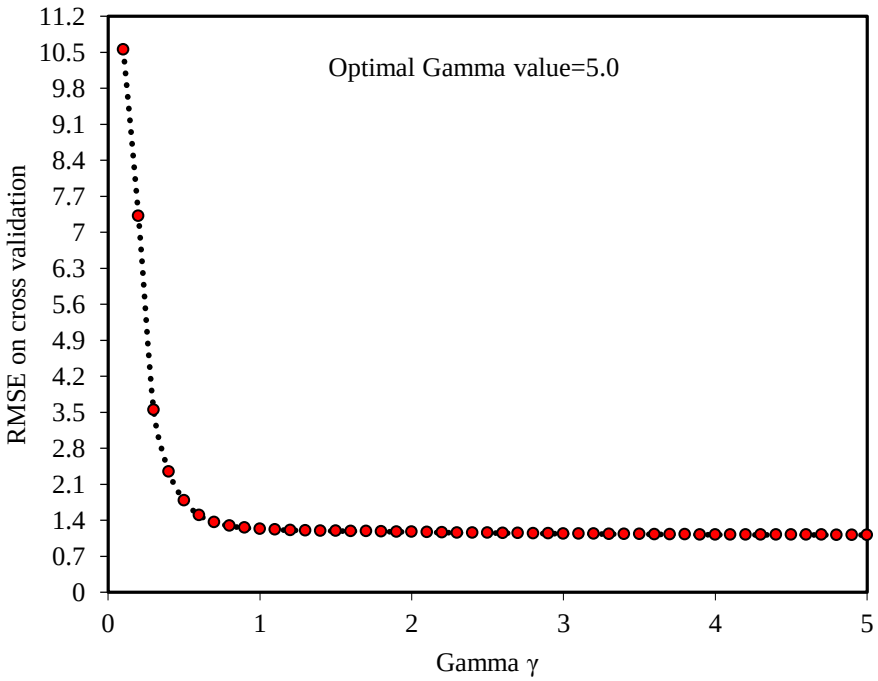
(A)


Optimal Ɛ value= 0.1
RMSE cross validation
Epsilon Ɛ
1.096
1.094
1.092
1.09
1.088
1.086
1.084
1.082
1.08
0
0.02
0.04
0.06
0.08
0.1
0.12

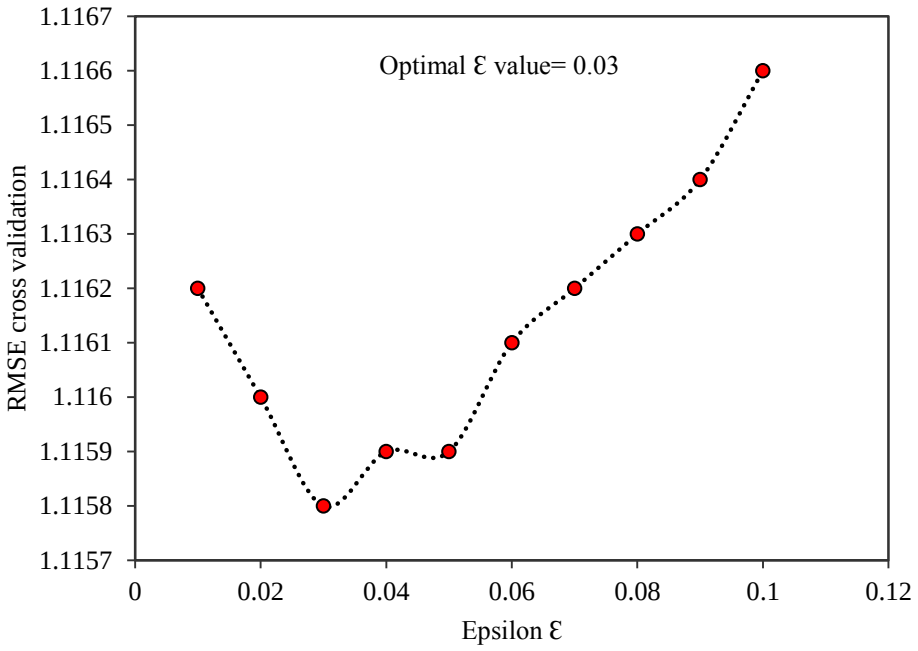
(B)


Optimal C value= 22
RMSE Cross Validation
Capacity parameter C
1.6
1.5
1.4
1.3
1.2
1.1
1
0
10
20
30
40
50

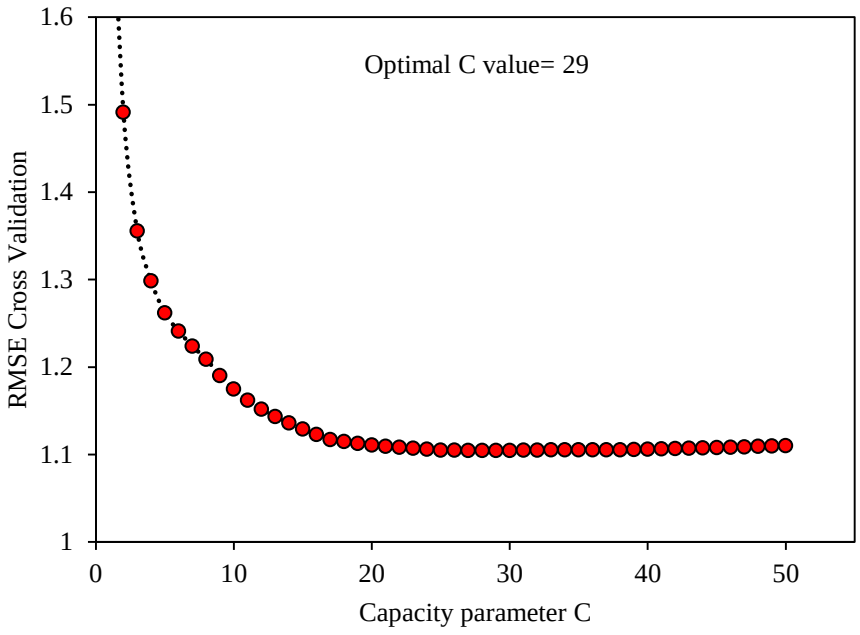
(C)


Optimal Gamma value=5.0
RMSE on cross validation
Gamma γ
11.2
10.5
9.8
9.1
8.4
7.7
7
6.3
5.6
4.9
4.2
3.5
2.8
2.1
1.4
0.7
0
0
1
2
3
4
5

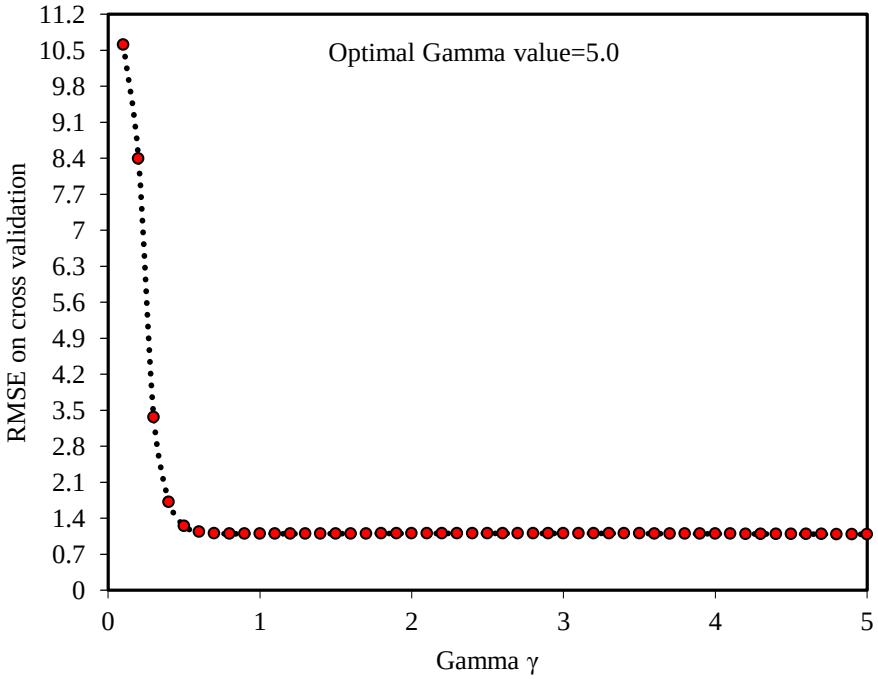
(D)


Optimal Ɛ value= 0.03
RMSE cross validation
1.1167
1.1166
1.1165
1.1164
1.1163
1.1162
1.1161
1.116
1.1159
1.1158
1.1157
0
0.02
0.04
0.06
0.08
0.1
0.12
Epsilon Ɛ

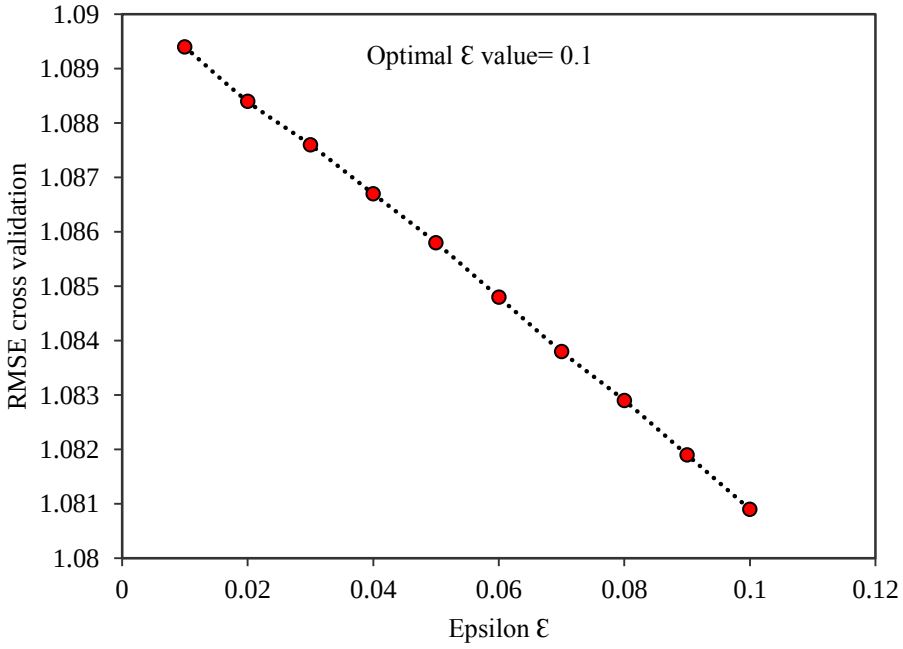
(E)


Optimal C value= 29
RMSE Cross Validation
1.6
1.5
1.4
1.3
1.2
1.1
1
0
10
20
30
40
50
Capacity parameter C

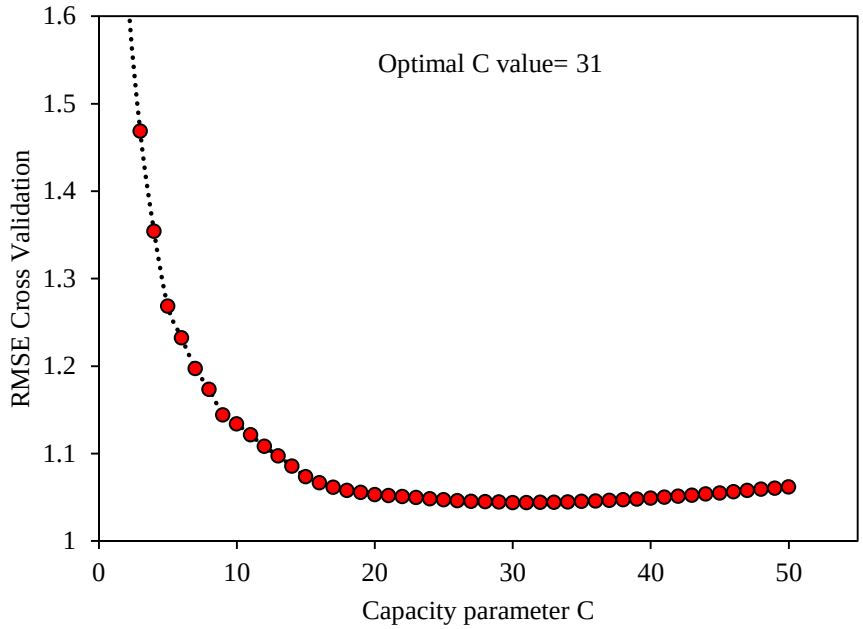
(F)


Optimal Gamma value=5.0
RMSE on cross validation
Gamma γ
11.2
10.5
9.8
9.1
8.4
7.7
7
6.3
5.6
4.9
4.2
3.5
2.8
2.1
1.4
0.7
0
0
1
2
3
4
5

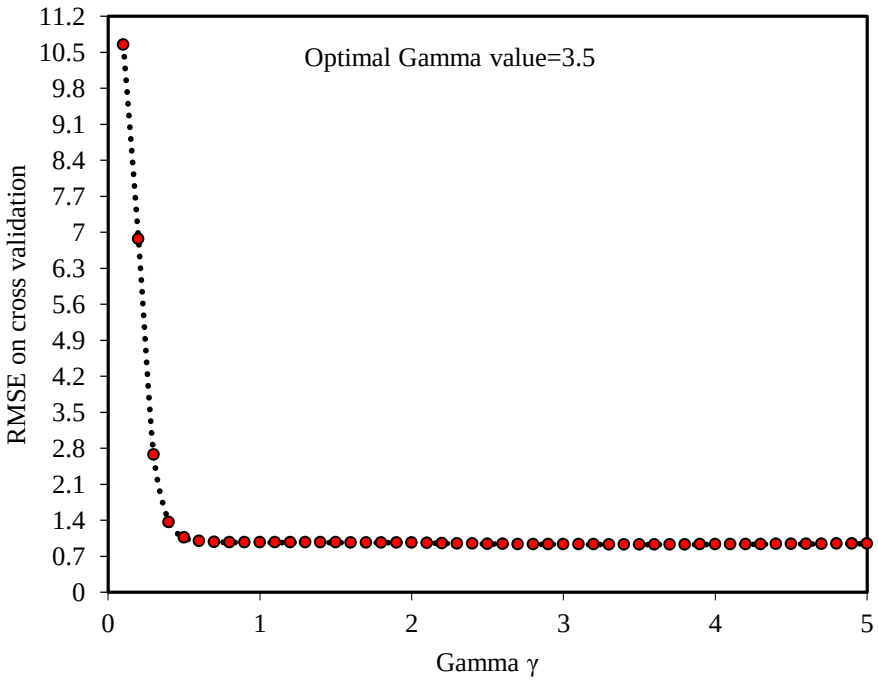
(G)


Optimal Ɛ value= 0.1
RMSE cross validation
Epsilon Ɛ
1.09
1.089
1.088
1.087
1.086
1.085
1.084
1.083
1.082
1.081
1.08
0
0.02
0.04
0.06
0.08
0.1
0.12

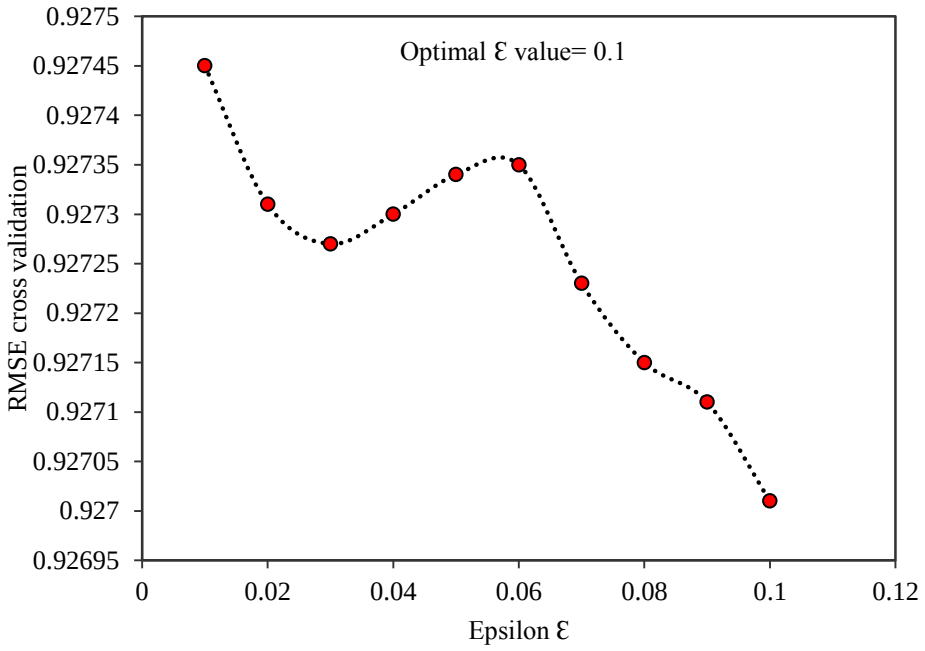
(H)


Optimal C value= 31
RMSE Cross Validation
Capacity parameter C
1.6
1.5
1.4
1.3
1.2
1.1
1
0
10
20
30
40
50

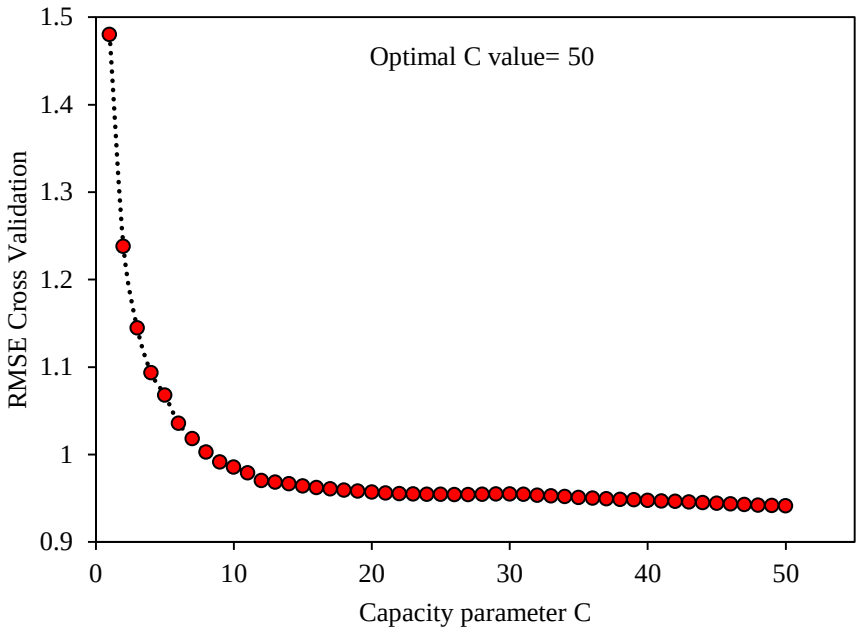
(I)


Optimal Gamma value=3.5
RMSE on cross validation
Gamma γ
11.2
10.5
9.8
9.1
8.4
7.7
7
6.3
5.6
4.9
4.2
3.5
2.8
2.1
1.4
0.7
0
0
1
2
3
4
5

(J)

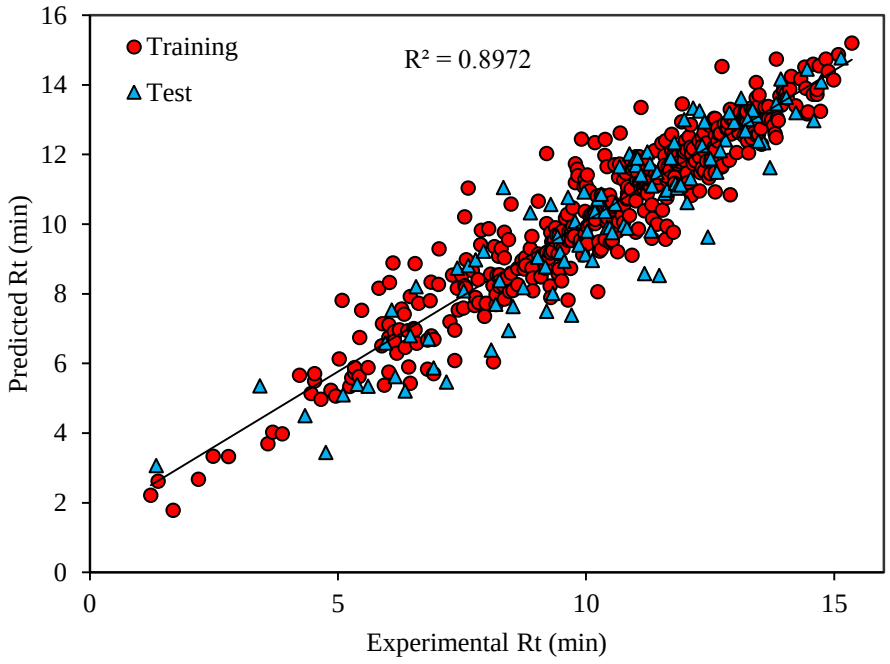

Optimal Ɛ value= 0.1
RMSE cross validation
Epsilon Ɛ
0.9275
0.92745
0.9274
0.92735
0.9273
0.92725
0.9272
0.92715
0.9271
0.92705
0.927
0.92695
0
0.02
0.04
0.06
0.08
0.1
0.12

(K)

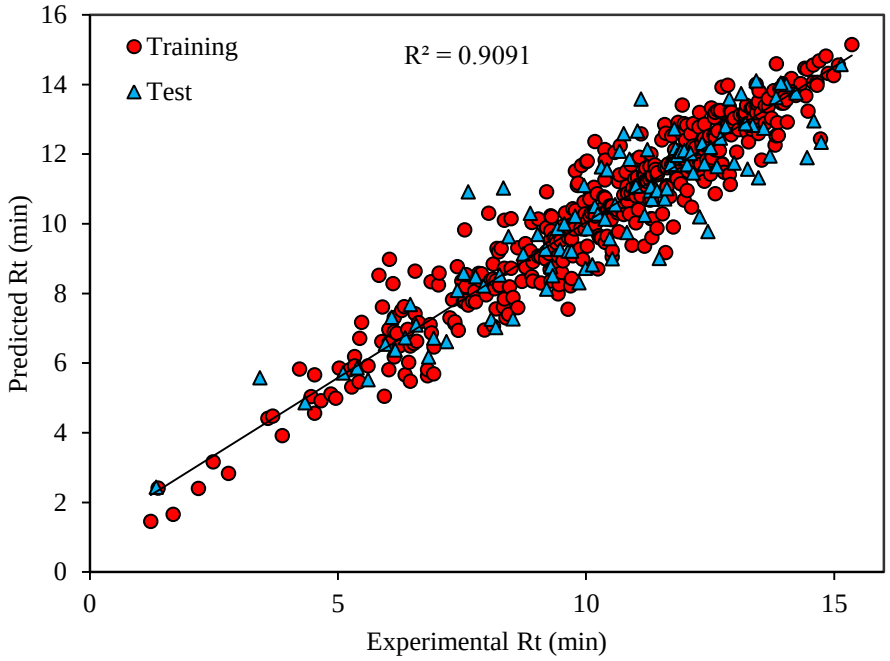

Optimal C value= 50
RMSE Cross Validation
Capacity parameter C
1.5
1.4
1.3
1.2
1.1
1
0.9
0
10
20
30
40
50

(L)

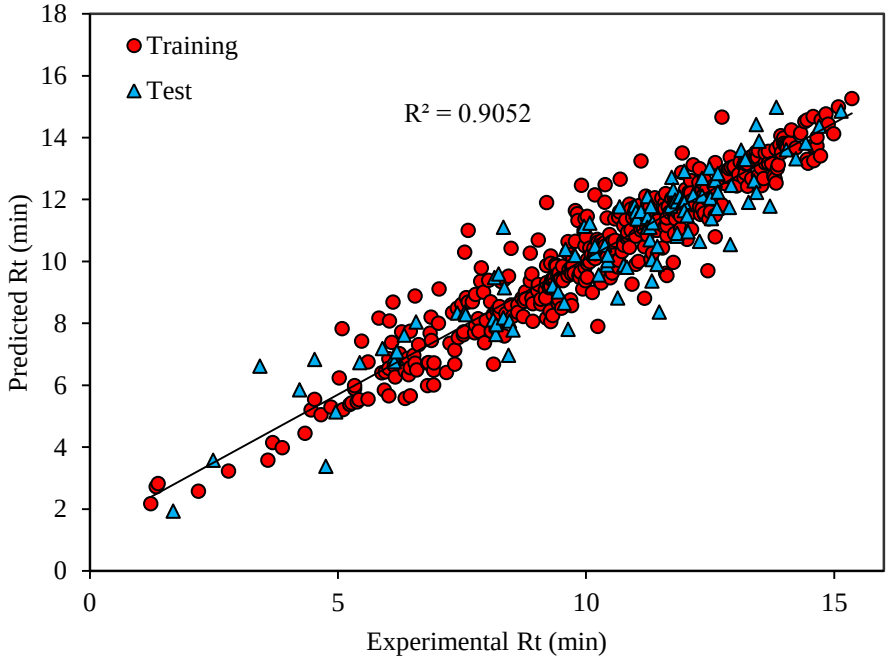
Figure S25. The SVM parameters vs. RMSE for the training set based on PCA-SW-SVM (A-C), PCA-GA-SVM (D-F), kNN-SW-SVM (G-I), kNN-GA-SVM (J-L) (positive ionization compounds)



(A)


Training
Test
R² = 0.9091
Predicted Rt (min)
Experimental Rt (min)
0
2
4
6
8
10
12
14
16
0
5
10
15

(B)


Training
Test
R² = 0.9052
Predicted Rt (min)
Experimental Rt (min)
0
2
4
6
8
10
12
14
16
18
0
5
10
15

(C)

Figure S26. The plot of predicted retention time against the observed retention time values based on (A) PCA-SW-SVM, (B) PCA-GA-SVM, and (C) kNN-SW-SVM.

kNN-GA-ANN

Since the models based on kNN and genetic algorithm showed appropriate internal and external results, for the generation of non-linear model based on ANN, kNN-genetic algorithm technique was considered. The selected compounds as valid and test set were marked in **Table S1** (positive ionization). The same newly introduced technique for choosing the nodes in negative ionization compounds were used here to develop accurate model without over-fitting problem. The results of this methodology were shown in **Table S21**. From Table S21, it can be seen that the model built based on node=6 shows the highest CCC value for the test, and also for the training set the calculated value is acceptable. Moreover, the calculated modified r^2 value for test set is the highest one among the other nodes. Considering the RMSE value for node 3 and 6, consequently the model based on node=6is being selected. The MPD values for training set with the different nodes were calculated using equation S15 and were given in **Table S21**. The obtained MPD values for node 3 and 6 indicates that model based on 6 nodes represents good fitting points. The predicted values based on kNN-GA-ANN were listed in **Table S1** and its strength as prediction tool were compared in **Table 4**. The predicted values for retention time by kNN-GA-ANN method for positive ionization compounds versus the observed retention time are shown in **Figure S27**.

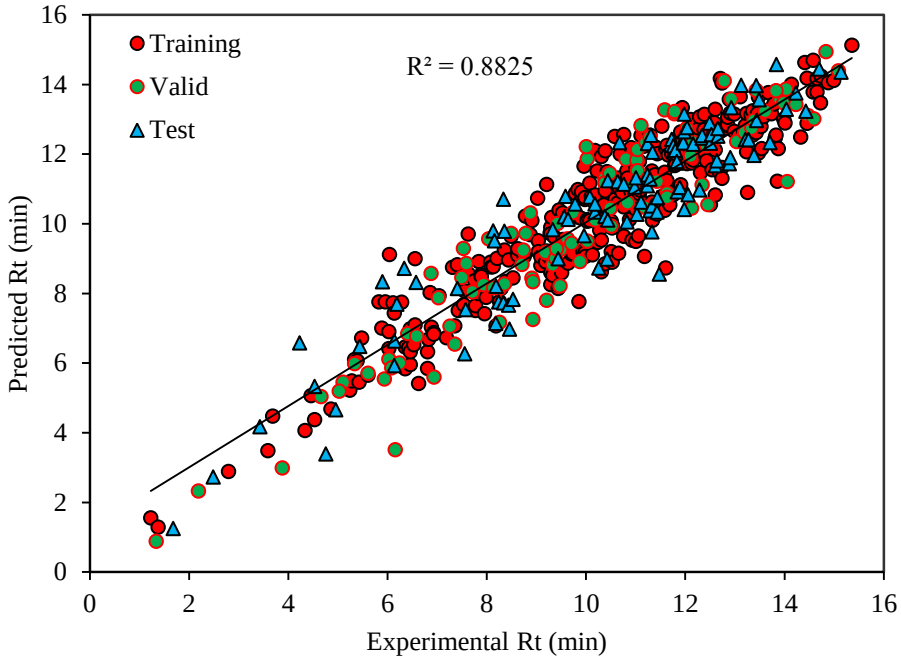


Figure S27. The plot of predicted retention time *versus* the observed retention time values based on kNN-GA-ANN (positive ionization)

SI 8. Supplementary content to Section 3: 5.2.2. Interpretation of descriptors

To understand BEHp2 effects and also its relationship with the polar atoms (O, N S and P), logD values, as well as charges potential and also BEHp2 values for some compounds were studied and listed in **Table S24**. Polar surface area (PSA) and charges potential were calculated based on DFT study on the basis of B3LYP/6-31*G method. From **Table S24**, m12 (Amitrole) showed the lowest BEHp2 in contrast to whole data set suggesting that atomic prolazability in substracture is very low. It seems that the presence of nitrogen in molecular structure decreases the BEHp2 values, in contrast to the presence of oxygen, sulfur and phosphorus. Comparing compound m12 with m219, it can be seen that the addition of an oxygen group inceased BEHp2 values but decreased logD values suggesting that the molecule is more polar, however as BEHp2 increased in compound m125, the logD is increased. Apart from the restriction of the compounds to interact with the stationary phase due to the lower value of this descriptor, BEHp2 values also seems to represent the steric effect of compounds over retention time indirectly due to the number

of fragments that can be derived from the core structure. A correlation analysis is carried out between molecular weight and BEHp2 values of the used compounds to seek for any relationship. The result of skipped correlation analysis is presented in Figure S28. There is a significant correlation between molecular weight and BEHp2 which explains why increasing BEHp2 values would increase the Rt.

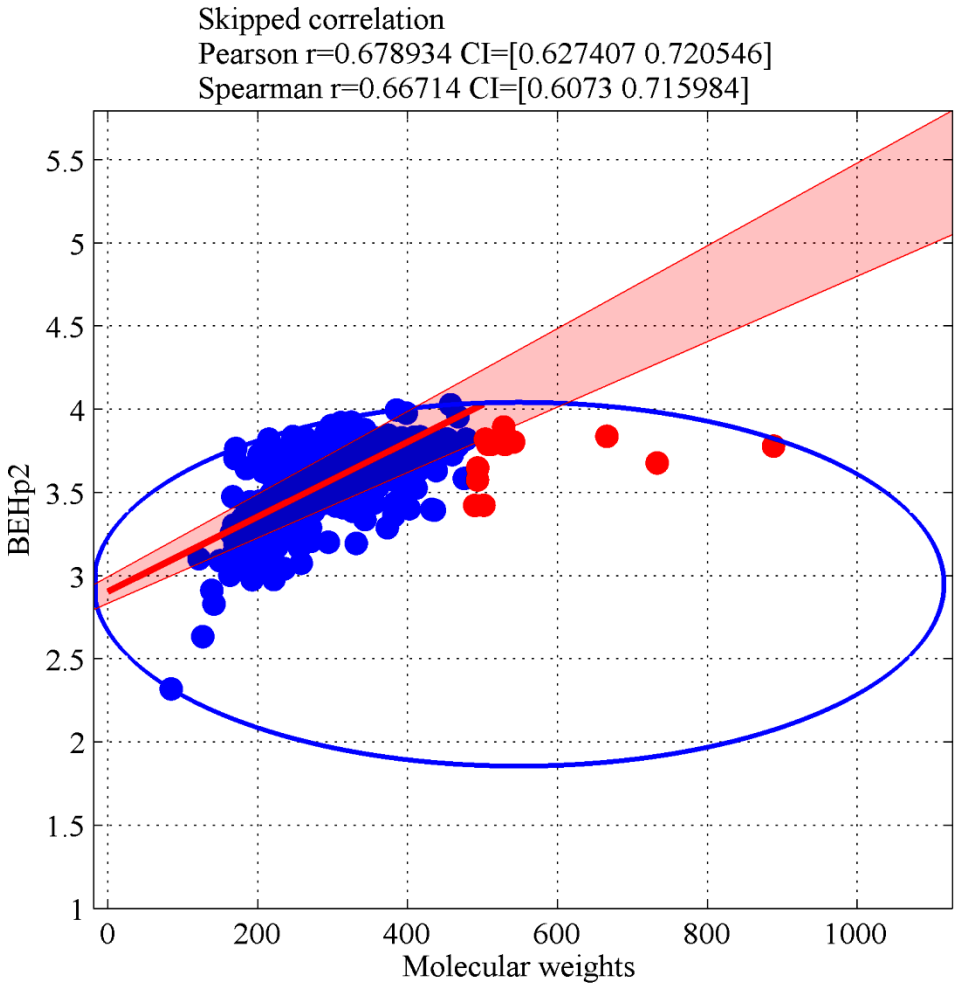


Figure S28. Skipped correlation analysis between BEHp2 and molecular weights

The density of molecular weights versus BEHp2 values is shown in **Figure S29**. As it can be seen, the major density of molecular weights is between 200-400 g mol^{-1} and, for this range of MWs, BEHp2 values are also increasing and show high density.

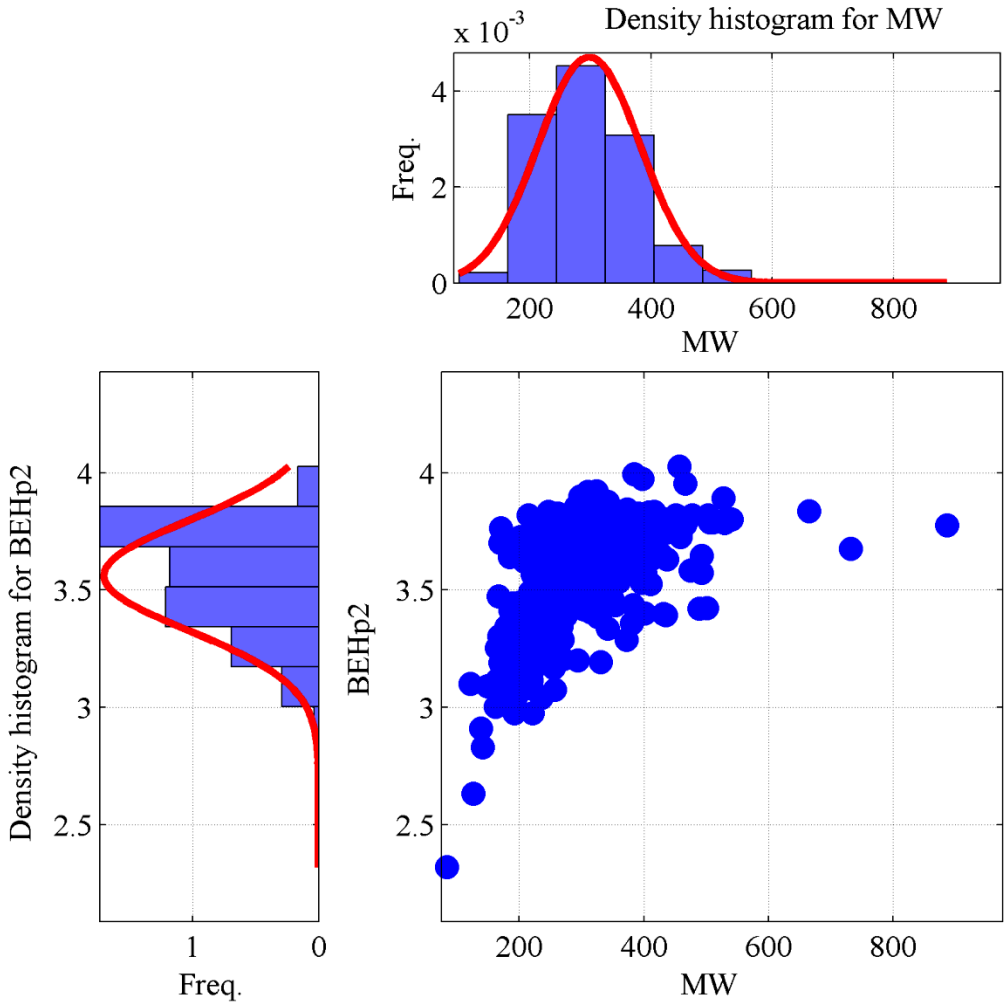


Figure S29. The density of molecular weights versus BEHp2 values

To check if BEHp2 shows any linear relationship with the retention time, component residual plots were derived. In this plot if the components fitted line shows high difference from residuals fitted line, it indicates that there is not any linear relationship between the predictor and dependent variable. This plot between BEHp2 and retention time is shown in **Figure S30**.

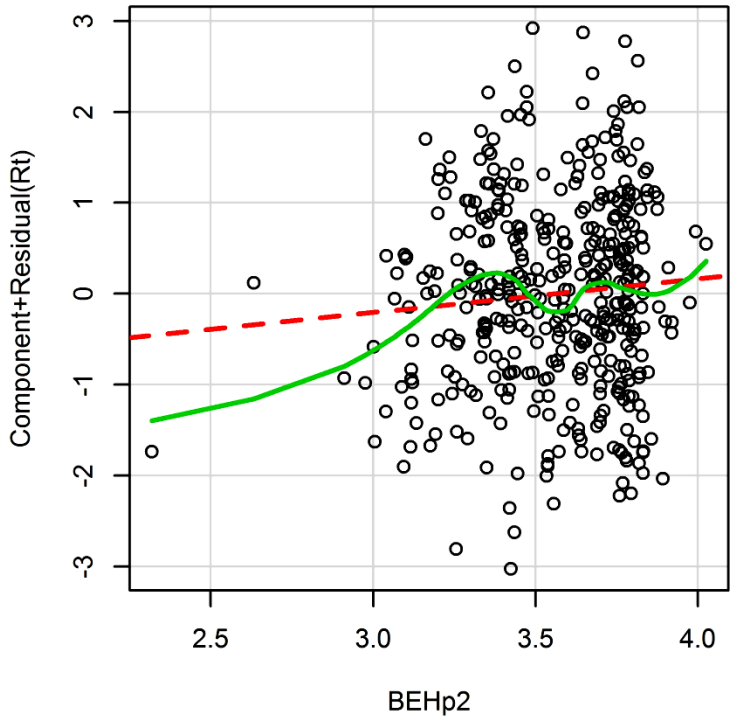


Figure S30. Component residual plots for BEHp2 and Rt

As it can be seen from **Figure S30**, there is not a robust linear relationship between this descriptor with Rt, suggesting that this descriptor is useful in interpretation of some specific compounds in dataset that the combination of effect of this descriptor with other ones would casue less prediction error.

There is also a masking effect of logD which its effect is more dominant than BEHp2. Therefore, it can be concluded that if a molecule represents more fragments with less number of nitrogen, the BEHp2 descriptor will increase, leading to an increase of the Rt. However, the masking effect of logD over Rt should be studied simultaneously.

SI 9. Instructions for using the OTrAMS

A MATLAB code has been developed for the visualization of the predicted retention times of suspect compounds and the interpretation of the outlier values. The code could be downloaded at the following link:

http://trams.chem.uoa.gr/docs/OTrAMS.p

Here are the instructions of using the code:

Using the developed models as filtering tool for LC-HRMS suspect screening, it could be understood if the detected compounds are within applicability domain of the models or not, by using OTrAMS. The following example provides instructions for using the developed "bubble plot".

First of all, the user should change the working directory of MATLAB software to the desired directory. Then, he/she prepares the descriptors list for the suspect compound as mentioned in section 2.5 (put the name of logD as "LogD" in representative column, since MATLAB is sensitive to capital letters; the names of compounds should be written in the first column in an excel file, the second column should contain the observed retention times, and the rest columns should contain the molecular descriptors. Then, the following command in MATLAB command window should be written:

OTrAMS(ESI, save, plot)

If the user would like to save the analysis results, she/he should put save=1, and if she/he would like to get the "bubble plot", she/he should put plot=2. Depending on their ESI mode, they can use ESI="Positive" or 'Negative'. If they would like to get the 2D-plot of the analysis, they should enter plot=1.

SI 10. Comparison of the developed models to the literature

The developed models were compared to the models which were already developed based on various evaluation criteria and can be found in **Table S25**.

References of SIF

(1) Abdi, H.; Williams, L. J., Principal Component Analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433-459.

(2) Jolliffe, I. T., *Principal Component Analysis*. 2nd ed. Springer: New York, 2002.

(3) Hartigan, J. A., *Clustering Algorithms*. John Wiley & Sons, Inc.: New York, 1975; p 351.

(4) Jain, A. K., Data clustering: 50 Years Beyond K-means. *Pattern Recogn. Lett.* **2010**, *31*, 651-666.

(5) Bodzioch, K.; Durand, A.; Kaliszan, R.; Bączek, T.; Vander Heyden, Y., Advanced QSRR Modeling of Peptides Behavior in RPLC. *Talanta* **2010**, *81*, 1711-1718.

(6) Hocking, R. R., A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression. *Biometrics* **1976**, *32*, 1-49.

(7) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; Jong, S. D.; Lewi, P. J.; Smeyers-Verbeke, J., *Handbook of Chemometrics and Qualimetrics, Part A*. Elsevier: Amsterdam, 1997.

(8) Draper, N. R.; Smith, H., *Applied Regression Analysis*. 2d Edition ed.; John Wiley & Sons, Inc: New York, 1981.

(9) Wagner, J. M.; Shimshak, D. G., Stepwise Selection of Variables in Data Envelopment Analysis: Procedures and Managerial Perspectives. *Eur. J. Oper. Res.* **2007**, *180*, 57-67.

(10) Leardi, R., Genetic Algorithms in Chemometrics and Chemistry: a review. *J. Chemometr.* **2001**, *15*, 559-569.

(11) Waller, C. L.; Bradley, M. P., Development and Validation of a Novel Variable Selection Technique with Application to Multidimensional Quantitative Structure−Activity Relationship Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 345-355.

(12) Goicoechea, H. C.; Olivieri, A. C., A New Family of Genetic Algorithms for Wavelength Interval Selection in Multivariate Analytical Spectroscopy. *J. Chemometr.* **2003**, *17*, 338-345.

(13) Ding, Q.; Small, G. W.; Arnold, M. A., Genetic Algorithm-Based Wavelength Selection for the Near-Infrared Determination of Glucose in Biological Matrixes: Initialization Strategies and Effects of Spectral Resolution. *Anal. Chem.* **1998**, *70*, 4472-4479.

(14) Aires-de-Sousa, J.; Hemmer, M. C.; Gasteiger, J., Prediction of 1H NMR Chemical Shifts Using Neural Networks. *Anal. Chem.* **2001**, *74*, 80-90.

(15) Leardi, R.; Boggia, R.; Terrile, M., Genetic Algorithms as a Strategy for Feature Selection. *J. Chemometr.* **1992**, *6*, 267-281.

(16) Goodarzi, M.; Heyden, Y. V.; Funar-Timofei, S., Towards Better Understanding of Feature-Selection or Reduction Ttechniques for Quantitative Structure–Activity Relationship Models. *TrAC Trends in Anal. Chem.* **2013**, *42*, 49-63.

(17) Lin, L., A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **1989**, *45*, 255-268.

(18) Mathworks Genetic Algorithm and Direct Search Toolbox Users Guide, The Mathworks Inc, 2005.

(19) Riahi, S.; Mousavi, M. F.; Shamsipur, M., Prediction of Selectivity Coefficients of a Theophylline-Selective Electrode Using MLR and ANN. *Talanta* **2006**, *69*, 736-740.

(20) Habibi-Yangjeh, A.; Pourbasheer, E.; Danandeh-Jenagharad, M., Prediction of Basicity Constants of Various Pyridines in Aqueous Solution Using a Principal Component-Genetic Algorithm-Artificial Neural Network. *Monatsh. Chem.* **2008**, *139*, 1423-1431.

(21) Dashtbozorgi, Z.; Golmohammadi, H.; Konoz, E., Support Vector Regression Based QSPR for the Prediction of Retention Time of Pesticide Residues in Gas Chromatography–Mass Spectroscopy. *Microchem. J.* **2013**, *106*, 51-60.

(22) Golmohammadi, H.; Dashtbozorgi, Z.; Acree Jr, W. E., Quantitative Structure–Activity Relationship Prediction of Blood-To-Brain Partitioning Behavior Using Support Vector Machine. *Eur. J. Pharm. Sci.* **2012**, *47*, 421-429.

(23) Tropsha, A., Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476-488.

(24) Golbraikh, A.; Tropsha, A., Beware of q2! *Journal of Molecular Graphics and Modelling* **2002**, *20*, 269-276.

(25) Vapnik, V., *Statistical learning theory*. Wiley: New York, 1998.

(26) Riahi, S.; Pourbasheer, E.; Ganjali, M. R.; Norouzi, P., Investigation of Different Linear and Nonlinear Chemometric Methods for Modeling of Retention Index of Essential Oil Components: Concerns to Support Vector Machine. *J. Hazard. Mater.* **2009**, *166*, 853-859.